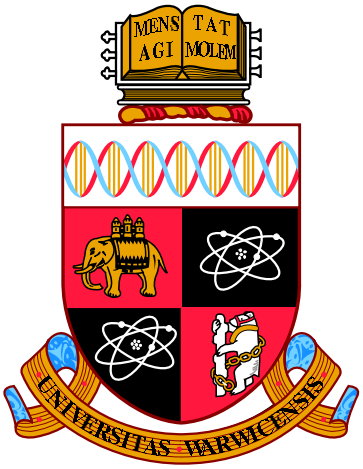

On the analysis of livestock networks and the modelling of foot-and-mouth disease

by

Peter Michael Dawson

Thesis

Submitted to the University of Warwick

for the degree of

Doctor of Philosophy

Centre for Complexity Science

November 2016

THE UNIVERSITY OF WARWICK

Contents

List of Tables

List of Figures

Acknowledgments

> "The man in black fled across the
> desert, and the gunslinger followed."
>
> —Stephen King, The Gunslinger[1978]

The research in this thesis was made possible by the Warwick Doctoral Training Centre for Complexity Science which was funded by the Engineering and Physical Sciences Research Council. The UK data was made available to is through Defra. The Turkish data set was provided by the Turkish Ministry of Food Agriculture and Livestock. The European Commission for the control of Foot-and-Mouth disease (EuFMD) facilitated our collaboration with the Turkish authorities. In particular we are grateful to Keith Sumption and Melissa McLaws of EuFMD and Naci Bulut of the Sap Institute in Ankara.

On a personal note I would like to thank both my supervisors Mike Tildesley and Matt Keeling for putting up with me for the past four years. I am very proud to say that I am the first of Mike's students to complete a PhD, I'm sure many more will follow me.

While PhD research is mostly a solitary affair I was very lucky to have great support and make many friends in Complexity throughout my MSc and PhD. In no particular order I'd like to thank my office mates: Davide Michieletto, Matt Graham, Tom Machon, Dario Papavassiliou, Neil Jenkins and Joel Nichols; the climbers: Davide and Stefi, Neil, Liz Buckingham-Jeffrey, Fede Botta, and Jonathan Skip; Tickle-Tackle: Davide, Neil, Matt, Anthony Woolcock, Ben St. Collyer, Ben Hu, Samik Datta, Ed Hill, and Chris Oates; disease buddies: Liz, Samik, Ed, Mike

Irvine, Dan Sprague, and Alex Bishop; Jamie WTF: Ben, Mike, Jamie Harris, Adam Miller, and Katy Calver. I'd also like to thank Ben and Samik for giving me bed and board during my write-up stage, I'm not sure if I wrote up any quicker but I got a lot better at Super Smash Bros.

I'd like to thank my parents for their support in everything I have ever done. Finally I'd like to thank Jen, I owe you more than I can say.

This thesis was typeset with LaTeX 2_ε[i] by the author.

[i] LaTeX 2_ε is an extension of LaTeX. LaTeX is a collection of macros for TeX. TeX is a trademark of the American Mathematical Society. The style package *warwickthesis* was used.

Declarations

Chapter 3 and its accompanying appendix (Appendix B) have been published as

- P. M. Dawson, M Werkman, E. Brooks-Pollock and M. J. Tildesley. 2015. Epidemic predictions in an imperfect world: modelling disease spread with partial data. *Proceedings of the Royal Society B* **282**:20150205.

Abstract

Foot-and-mouth disease (FMD) is a highly infectious disease affecting cloven-hoofed ruminants. FMD is endemic across Asia, Africa and South America and, as demonstrated by the 2001 outbreak in the UK, can cause devastating epidemics in FMD free countries.

A comprehensive dataset comprising the births, deaths and farm-to-farm movements of almost fifty million cattle was made available to us by the Turkish authorities. In chapter 2 we discuss the processing, cleaning and analysis of this dataset. Cattle movement networks were built using farms as nodes and the farm-to-farm movements as edges before performing a network analysis.

The richness and completeness of the Turkish data set is not generally available to modellers. In chapter 3 we implement various sampling strategies to determine the quantity of network data required to give accurate epidemiological predictions, using the British cattle trade network as a case study.

In chapter 4 we introduce community detection as a method for analysing the cattle movement networks of both the UK and Turkey. This analysis goes beyond the calculation of the farm level statistics calculated in chapter 2 and seeks to find higher order structures in these networks.

Chapter 5 develops a within farm model of FMD by drawing on outbreak data from two farms of different sizes in Turkey. This model is parametrised by running a Bayesian inference scheme on the smaller of the two farms. Simulations were then performed on both farms using the inferred posterior parameter distributions.

Abbreviations

ABC	Approximate Bayesian computation
AMLS	Animal movement licence scheme
APHA	Animal plant health agency
BSE	Bovine spongiform encephalopathy
BTB	Bovine tuberculosis
CP	Contiguous premise
CTS	Cattle tracing system
DC	Dangerous contact
Defra	Department for environment, food and rural affairs
FAO	The Food and Agriculture Organization of the United Nations
FMD	Foot-and-mouth disease
FPR	False positive rate
GSCC	Giant strongly connected component
MDI	Maternally derived immunity
MCMC	Markov chain Monte-Carlo
NASS	National agricultural statistics service

NG	Newman-Girvan null-model for community detection
LR_+	Positive likelihood ratio
NSP	Non-structural proteins
OIE	The World Organization for Animal Health
PRC	Particle rejection control
RMS	Random movement sampling
RNS	Random node sampling
ROC	Receiver operator characteristics
SBS	Snowball sampling
SMC	Sequential Monte-Carlo
SP	Structural proteins
TNS	Targeted node sampling
TPR	True positive rate
VE	Vaccine effectiveness

Chapter 1

Foot-and-Mouth disease and the modelling of infectious disease

> I recollect a singular contagion of the year 1514 which affected only oxen; it first showed itself in the district of Frioli, then gradually extended to northern Italy and thence reached our country. At first the ox went off its feed without any obvious cause, but if the herdsmen looked into the mouth a certain roughness and small pustules were noted on the palate and the whole oral cavity. It was necessary at once to isolate the infected beast from the rest of the herd, otherwise all became affected. Gradually the disease descended to the shoulders and thence to the feet, and almost all the beasts in which this happened recovered, but when it did not, they usually died.
>
> Hieronymus Fracastorius [1546]

Foot-and-mouth disease (FMD) is a highly infectious, viral disease which affects cloven-hoofed ruminants. Primarily cattle, buffalo, sheep, goats and pigs are affected but antelope, deer and bison have also been infected[Arzt et al., 2011a,b]. The disease manifests as blisters on mouth, feet, nose and udders of the infected animal.

In regions of the world where FMD is endemic there are large numbers of cattle which act as a reservoir and though small ruminants can play an important role in the transmission process it is not known if FMD can be maintained without a large cattle population [OIE, 2012].

There are seven serotypes of FMD and it is generally accepted that infection with one serotype does not confer immunity against another [Mardones et al., 2010], though there is no evidence to suggest that an animal can be infected by two

serotypes at the same time. Each of these serotypes is further divided into multiple strains.

In this thesis we examine the techniques for predicting the size and extent of FMD epidemics in the UK and build a model to aid in our understanding of endemic FMD in Turkey. This introduction aims to put FMD in context and introduces immunity and vaccination of which a basic understanding is necessary before embarking on the development of mathematical models.

1.1 FMD in a global context

Hieronymus Fracastorius described an infectious disease of cattle with symptoms resembling the disease we know as FMD. Fracastorius' account was dated 1514 suggesting that FMD is at least 500 years old. We shall describe the current state of FMD. Though a detailed and fully referenced history of FMD during the 20th century is available in Jamal and Belsham [2013][i].

Member nations of the OIE are given an FMD status which is updated on the OIE website [OIE, 2015], figure 1.1. The entire EU and Kazakhstan, North America, Oceania, Japan and regions of both South America and the south of Africa are currently classified as free from FMD without vaccination. The majority of the South American continent is FMD free with vaccination while Africa and Asia do not have an official status for FMD. FMD is endemic across Asia and most of Africa.

In general FMD is not fatal and most animals will recover though they may suffer from reduced milk yield and lameness [Grubman and Baxt, 2004]. However due to the efforts to contain the disease and international trade bans on countries without FMD free without vaccination status the economic costs of the disease are estimated at between US$6.5 million and US$20 million a year. These large economic losses due to endemic FMD tend to hit the world's developing countries the hardest where more people are directly dependent on livestock. A breakdown of the effects of FMD and estimates of their associated costs is given in Knight-Jones and Rushton [2013].

The impact of FMD is not limited to developing countries where the disease is endemic. Over the course of the past 15 years there have been several incursions of FMD into traditionally disease free countries which have had devastating consequences to the agricultural sectors and the wider economies of the affected countries. The most widely documented of these was the 2001 outbreak in the UK [Anderson, 2002]. During this outbreak a total of 10 million cattle and sheep were

[i]For a pre-20th century picture try Greenwood [1927] which comes with a map.

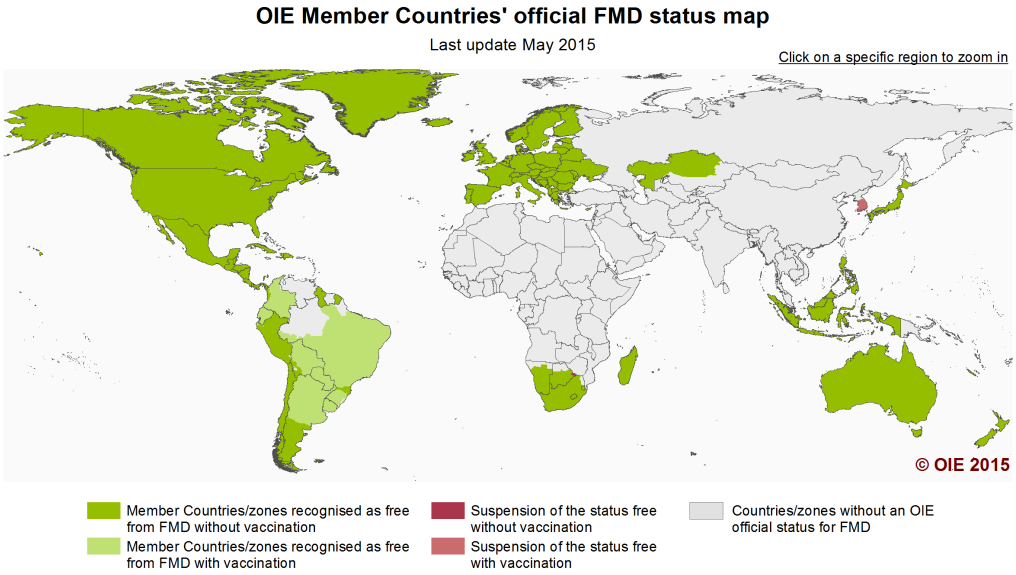

Figure 1.1: OIE member countries' offical FMD status map last updated in May 2015. This map is reproduced from the OIE website.

culled and the crisis was estimated to have cost the UK US$16 billion [Anderson, 2002; National Audit Office, 2002]. Other notable epidemics have occured in the Miyazaki prefecture of Japan [Nishiura and Omori, 2010] and in South Korea both in 2010 [Yoon et al., 2015].

Whilst FMD is a global disease not all of the seven serotypes are ubiquitous (figure 1.2). Serotype O is the most common and the most studied - it was a strain of O that caused the UK 2001 outbreak - and is seen in all regions suffering from FMD. Described as pandemic, serotype O is designated as the PanAsia strain which was confined to India until the late 1990s. Since 1996, serotype O has spread across most of Asia, through the Middle East and into Africa and South America. A comprehensive list of the dates of outbreaks of serotype O across the world is available in Jamal and Belsham [2013] with more detailed discussion and classification of regions that have been infected by different strains of O available in Rweyemamu et al. [2008].

The next most common serotype is A followed by Asia-1 which has not yet been seen in Africa or South America. There are three strains dubbed SAT-1, SAT-2 and SAT-3 which have emerged from the Southern African Territories and have not been seen outside of Africa. The SAT strains differ from the pan-Asian strains in that they do not appear to rely on a reservoir of domestic cattle to persist and can be maintained in African buffalo. The seventh serotype, C, has had not been

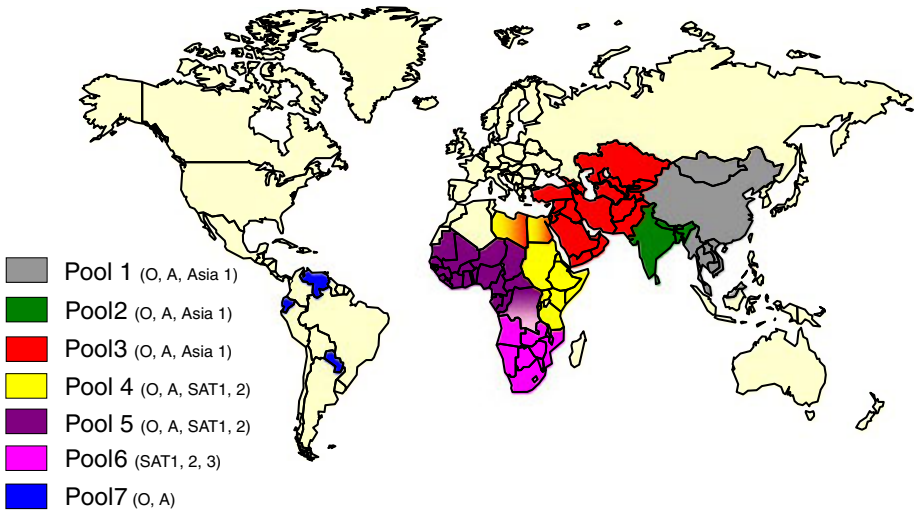

Figure 1.2: Geographical distribution of seven pools of foot-and mouth disease viruses. Serotype O FMDV is the most widely distributed serotype of the virus (in 6 of the 7 indicated virus pools) whereas, in contrast, SAT3 is only present in pool 6 (within southern Africa). The Asia-1, SAT1 and SAT2 serotypes also have quite limited geographical distribution. However, individual countries can have multiple serotypes in circulation at the same time and hence it is necessary to be able to determine which serotype is responsible for an outbreak if vaccination is to be used. Countries which are normally free of the disease (marked in yellow) can still suffer incursions of the virus which can have high economic costs. Reproduced from Jamal and Belsham [2013]

reported since 2005 in Ethiopia and may now only exists in laboratories [Jamal and Belsham, 2013].

1.2 Transmission and diagnosis

Outbreaks of FMD are normally associated with the movement of infected animals (either domestic or wild) and their subsequent contact with susceptible animals or through indirect transmission of contaminated products. Airborne transmission has also been suggested given the correct environmental conditions [Grubman and Baxt, 2004]. As cattle are believed to be the main drivers of the disease and as the data we have available from both Turkey and the UK only pertains to cattle we shall focus on the disease process in cattle only.

After being exposed to FMD virus a cow enters a latent period during which the cow is infected but is not infectious [Arzt et al., 2011b]. Latent is often synonymous with incubation but for FMD there is a difference. The incubation period generally refers to the time from exposure to symptomatic or clinical, while a cow

infected with FMD may become infectious sub-clinically [Charleston et al., 2011].

All secretions and excretions of an infectious cow may contain FMD virus, including respired air. After acute infection cattle recover though in some cattle the virus may persist these cattle are descriped as being in a carrier state. There is no experimental evidence that cattle in a carrier state can transmit the virus to naïve cattle [Grubman and Baxt, 2004].

Animals are first diagnosed clinically; symptoms include a high fever and excessive salivation and the formation of vesicules mainly in the mouth and between the digits of the feet. These symptoms are not just signs of FMD and can also be brought about by other diseases such as vesicular stomatisis. Laboratory tests are therefore also necessary to correctly diagnose FMD and to classify the serotype and strain of the virus.

1.3 Vaccination and immunity

Control of infectious disease both in humans and animals has been practised for centuries. The goal of any control measure is to to break the transmission chain between the infected host and the rest of the susceptible population. There are two obvious ways to achieve this; either remove the infected host or directly protect the population.

Removing the infected host can be achieved through isolation in the form of quarantine. The word quarantine comes from the Venetian dialect *quranta giorni* - forty days, and relates to the forty days ships and vistors to the city of Dubrovnik has to spend in isolation for fear of the Black Death in the 14^{th} century. The practice of quarantine itself dates further back to at least biblical times, Leviticus 14-46:

"*All the days wherein the plague shall be in him he shall be defiled; he is unclean: he shall dwell alone; without the camp shall his habitation be*"

In the veterinary world quarantine is not always feasible and removal of the infected host is often carried out by culling of not only the identified infected host but also the host's contacts whose disease status may be unknown. In the 2001 UK outbreak this strategy of culling not only the infected host's farm but also culling contiguous premises (CPs) and other farms identified as dangerous contacts (DCs) was the adopted [Tildesley et al., 2009], and ultimately successful, control strategy, though one that was and remains highly controversial [Kitching et al., 2005]. The costs of culling however are very high, both economically with compensation being paid to affected farms, and politically as piles of burning, slaughtered cattle conjure an emotive image.

While culling can be highly effective in the containment of an epidemic of FMD it is generally politically infeasible in a country such as Turkey, where FMD is endemic. In the endemic situation where the goal is to lower the disease burden until eventual eradication, control must be practised through improved animal health practices and mass vaccination campaigns.

A global strategy for the control of FMD has been composed jointly by the OIE and FAO [FAO and OIE, 2012]. The FMD Progressive-Control-Pathway (PCP) is a key feature of this strategy and acknowledges that vaccination will be the main control strategy [Sumption et al., 2012].

1.3.1 The mechanism of vaccination

The FMD virus enters infects healthy cells in the animal, it then replicates. During replication the FMDV linear RNA is used to create structural proteins (SP) and non-structural proteins (NSP). The NSP aid production of SP and when the infectious virus is produced within an infected cell NSP are left in the cell [Doel, 2003; IAEA and FAO, 2007], for a schematic see figure 1.3.

Vaccine is produced in a laboratory by the inactivation of the FMD virus which is then purified of NSP. Thus, if a cow has been injected with a well purified vaccine it will test positive for antibodies to SP but negative to antibodies of NSP. A cow will only test positive to NSP if it has acquired infection naturally and is actually in a recovered, not vaccinated stated. Recent advances in the production of vaccines have therefore allowed for the development of a test which distinguishes between cattle that have been naturally infected with FMD and those that have been vaccinated [IAEA and FAO, 2007; Paton et al., 2014]. This test is of vital importance to countries seeking to gain or regain the moniker 'free from FMD with vaccination' from the OIE.

Vaccines are serotype and strain specific and must be manufactured based on what form of FMD is prevalent in a region. It is possible to vaccinate against multiple serotypes simultaneously by producing polyvalent vaccines [Doel, 2005; Knight-Jones, 2014].

Vaccines have an associated efficacy which is the reduction in a attack of vaccinated individuals compared to unvaccinated individuals, under ideal conditions. This is tested for in clinical trials. However as field work does not take place under ideal conditions vaccine effectiveness (VE) must be calculated instead [Knight-Jones et al., 2014b]. VE is associated with how well the vaccine matches the circulating strain of FMD, the receiver animal's vaccination history, storage of the vaccine, adherence to the cold-chain, purification and delivery of the vaccine, and the vaccine's

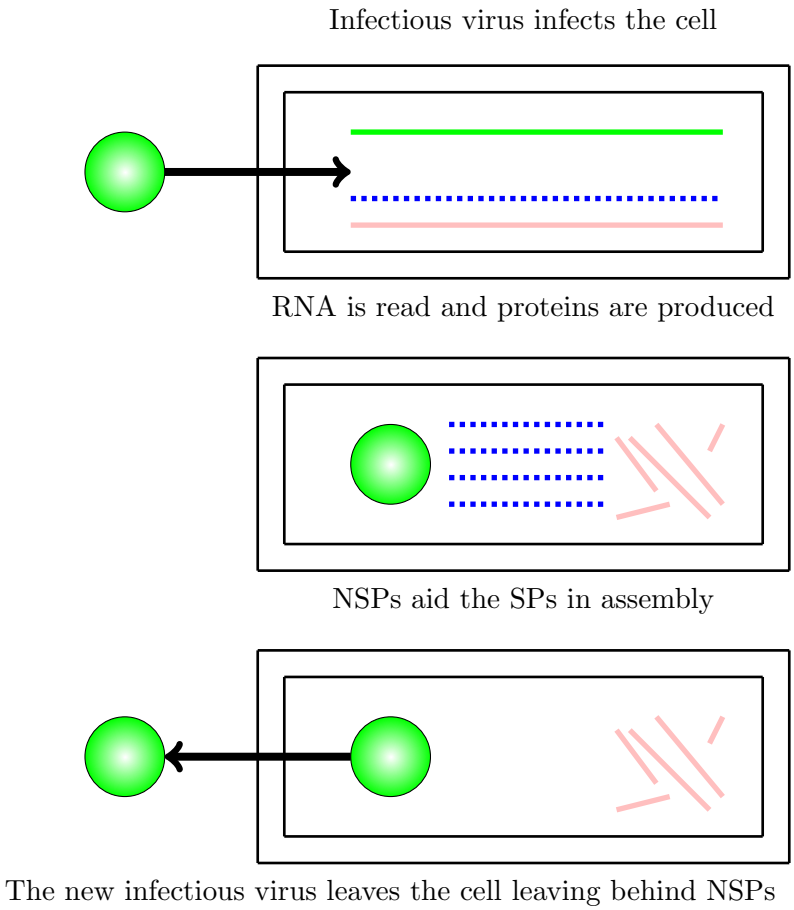

Figure 1.3: Sequence of events in a FMD infected cell. A healthy cell is infected by FMD virus, RNA is read and both structural (blue) and non-structural (pink) proteins are produced. The non-structural proteins (NSPs) aid the structural proteins (SPs) in assembling new FMD virus. The new infectious virus leaves the cell but NSPs are left behind. Figure inspired by IAEA and FAO [2007].

efficacy.

Due to these factors the vaccine effectiveness must be determined by field studies. One such field study, carried out in Turkey, comprised the PhD work of Theo Knight-Jones [Knight-Jones, 2014; Knight-Jones et al., 2014a] in which he determined that the FMD Asia-1 TUR 11 vaccine had a VE of 69% [95% CI: 50% to 81%].

It is recommended that calves be vaccinated between 2-4 months and then again one month later [Doel, 2003], this dose course is prescribed to ensure successful vaccination if the first dose were to fail due to maternally derived immunity (see 1.3.2) . Protection from a course of vaccination typically last 6 months [Doel, 2005], as such mass vaccination campaigns in Turkey are generally carried out twice a year [Knight-Jones, 2014].

1.3.2 Maternally derived immunity

A calf whose mother has sufficient antibodies to FMD virus will derive some level of immunity to the virus. This maternally derived immunity (MDI) to FMD virus has been reported to persist in calves for up to 5 months [Kitching and Salt, 1995], though this will of course be strain dependent.

MDI will successfully prevent calves from become infected with FMD. Unfortunately these antibodies can also prevent the calf being successfully vaccinated. MDI levels can wane sufficiently so that that calf can become infected by FMD while still being high enough to prevent the calf developing an adequate immune response to vaccination [Doel, 2005; Kitching and Salt, 1995].

Indeed due to loss of MDI before successful vaccination in FMD outbreaks in both Turkey [Knight-Jones, 2014] and Bolivia [Gonzales et al., 2014] the age at which cows were most at risk to FMD was calculated as between 6-18 months.

Humans and other primates receive their mother's antibodies through the placenta. However, this is not the case for calves. Calves receive antibodies through colostrum which is contained in their mother's milk. As such, depriving a calf of its mothers milk will prevent the calf from gaining the maternally derived antibodies and hence from becoming immune, allowing for vaccination at a very young age [Roeder and Taylor, 2007].

1.3.3 Infection induced immunity

For some diseases after an individual has ceased to be infectious they may immediately be susceptible to reinfection, for other diseases once an individual has ceased

being infected they can never be reinfected and are said to be recovered, having gained immunity to the disease. However for some diseases this immunity may wane, individuals will eventually cease to be protected and may be reinfected.

It is not clear how immunity wanes in cattle previously infected with FMD. Doel [2005] reports that cattle reinfected with a homologous strain of FMD remained immune to reinfection six months after the initial infection, while other earlier studies report some cattle remaining immune for 4.5 - 5.5 years after infection. A recent study by Pomeroy et al. [2015] in Cameroon found that immunity to serotype O lasted on average 3.8 years while for serotype A they concluded that immunity was lifelong.

1.4 Infectious disease modelling

The study of infectious diseases through mathematical modelling lays its foundations in the early work by Sir Ronald Ross Ross [1911] an Indian born British medical doctor who won the Nobel prize for showing that mosquitoes are the host vector for malaria. Ross [1911] showed that instead of having to eradicate all mosquitoes in order to eradicate malaria it would enough to reduce their number below a certain threshold.

Disease modelling became more mathematically rigorous with the introduction of an ordinary differential equation (ODE) model by Kermack and McKendrick [1927]. Interest in infectious disease modelling waned slightly during the middle part of the 20^{th} century but was revitalised by several authors most notably by Roy Anderson and Robert May [Anderson and May, 1979a,b]. Most work during the last two decades of the 20^{th} century focused on theory with many influential papers and books being published, [Anderson and May, 1991; Dietz, 1993]. While theory is still being developed the increased availability of computer power has led much research to be focused on large, stochastic simulations rather than the solving of deterministic equations.

One quantity of particular interest in the study of infectious disease is the basic reproductive number denoted as R_0 (for a history of R_0 see Heesterbeek [2002]. This is defined as the average number of secondary cases arising from one infected individual in an entirely infectious population. It is clear that if $R_0 < 1$ the disease must die out. A great deal of effort is invested in the calculation of R_0.

The most typical model of infectious disease is a compartmental model. Individuals are placed in a compartment corresponding to their disease status and they progress through the compartments according to some rates. The SI model is

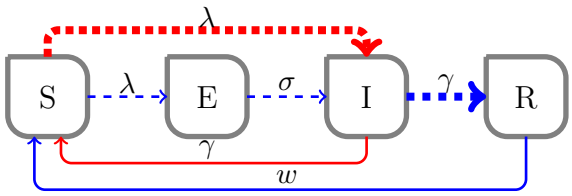

Figure 1.4: Depiction of an assortment of compartmental models for disease spread. Transitions between classes in each model are shown by different styles of line. The SI model by dotted lines, the SIS model by red lines, the SIR model by thick lines, the SEIR model by dashed lines, and the SEIRS by blue lines. The rate at which each transition occurs is given above the line.

the most simple of these models; individuals are susceptible (S) until they become infectious (I) with some rate λ. If we allow for individuals to recover (or die) from the disease then the model becomes SIR with R indicating the recovered state. Alternatively we can allow individuals to immediately become susceptible again, SIS. The length of time an individual spends on average in the I state is known as the infectious period, T. Individuals will cease to be infectious at some rate $\gamma = 1/T$.

The basic models can easily be made more complex depending on the disease being studied. For many diseases there is a latent period during which the individual has been infected but is not yet infectious. This class is donated by E for exposed and creates the SEIR model. Individuals will move from exposed to infectious at a rate σ.

Waning immunity can also be considered by allowing recovered individuals to become susceptible again, which they will do at a rate w. Incorporating waning immunity creates the SIRS or the SEIRS model. Figure 1.4 graphically depicts the models discussed so far.

The rate of infection λ is typically a function of the number of infected individuals in the population while the other rates γ, σ, and w are independent.

1.4.1 ODE models

Once a model has been decided on it is possible to write down a system of ordinary differential equations (ODEs) that describe the rate of change of the proportion of individuals in each compartment. The system of ODEs for the SIR model is given

below and was first written down by Kermack and McKendrick [1927]

$$\frac{\mathrm{d}S}{\mathrm{d}t} = -\beta IS \tag{1.1}$$
$$\frac{\mathrm{d}I}{\mathrm{d}t} = \beta IS - \gamma I \tag{1.2}$$
$$\frac{\mathrm{d}R}{\mathrm{d}t} = \gamma I \tag{1.3}$$

and we enforce that $S+I+R = 1$ so that the total size of the population. Examining equation 1.2 we note that if at the start of the epidemic the fraction of susceptible individuals $S(t = 0)$ is less than γ/β then $\frac{\mathrm{d}I}{\mathrm{d}t} < 0$ and the epidemic will die-out.

If on the other hand $S(t = 0) > \gamma/\beta$ then the rate of change of proportion of infected individuals is positive so the epidemic will take-off. The ratio of the rate of infection to the rate of recovery therefore defines the the basic reproduction number

$$R_0 = \frac{\beta}{\gamma}. \tag{1.4}$$

In the form considered above the SIR equations have assumed the population is closed, there are no births or natural deaths, this assumption is valid if the time-scale of the epidemic if much less than that of a natural lifespan. However, in the case of endemic disease we must include demography to allow for a changing population such that:

$$\frac{\mathrm{d}S}{\mathrm{d}t} = \nu - (\beta I - \mu S) \tag{1.5}$$
$$\frac{\mathrm{d}I}{\mathrm{d}t} = \beta IS - (\gamma + \mu)I \tag{1.6}$$
$$\frac{\mathrm{d}R}{\mathrm{d}t} = \gamma I - \mu R \tag{1.7}$$

where we have ν is the birth rate and μ death rate. We can once again derive an expression for R_0 with the death rate effectively acting to increase the rate of recovery and decrease R_0,

$$R_0 = \frac{\beta}{\gamma + \mu}. \tag{1.8}$$

Now that we have a population that changes through time we can calculate the criteria for the diseases becoming endemic, by searching for fixed points (S^*, I^*, R^*). Setting 1.6 to zero we obtain,

$$I^*(\beta S^* - (\gamma + \mu)) = 0 \tag{1.9}$$

which has two solutions, $I^* = 0$ which is the disease free state or

$$S^* = \frac{\gamma + \mu}{\beta} = \frac{1}{R_0} \tag{1.10}$$

implying the proportion of susceptible individuals required for a disease to become endemic is inveresly proportional to R_0. The second fixed point for I^* is

$$I^* = \frac{\mu}{\beta}(R_0 - 1), \tag{1.11}$$

since $I^* \geq 0$, R_0 again acts as a threshold condition that in order to have an endemic state R_0 must be greater than one.

With the importance of the basic reproductive number established we now seek ways to control the disease by effectively reducing R_0 to be below one.

1.4.2 Herd immunity

In order to eradicate a disease it is not necessary to vaccinate the entire susceptible population, it is enough to vaccinate a proportion of the the population, p. This can be easily shown by editing the birth rate in equation 1.5 to be

$$\nu' = \nu(1 - p) \tag{1.12}$$

which is balanced by a including a νp term in equation 1.7. After a change of variables we can solve the equations to write down a modified basic reproduction ratio

$$R_0' = (1 - p)R_0. \tag{1.13}$$

Imposing $R_0' < 1$ gives the the threshold proportion of the population that must be vaccinated at birth to eradicate the disease

$$p_c = 1 - \frac{1}{R_0}. \tag{1.14}$$

This phenomenon which allows for the theoretical eradication of disease is known as herd immunity and a review of its standing in the literature is given by Fine [1993].

This calculation for p_c comes with some caveats and assumptions. While vaccinating p_c of the population will lead to eradication, the time-frame for this eradication is not explicit, it may take many generations for herd immunity to be realised. A perfect vaccine which does not wane with time has been implicitly

assumed, while some vaccines do confer lifetime immunity many do not and hence the entire population may have to be constantly vaccinated to prevent resurgence of the disease.

The estimate for R_0 also has to be accurate, a slight underestimation can lead to a drastic increase in the numbers that must be vaccinated. So far only two diseases rinderpest in cattle [Roeder and Taylor, 2007] and smallpox in humans have been eradicated, both of these eradications required coordinated vaccination programmes.

For FMD vaccination can be used in two ways, either reactively in the case of an outbreak, or as a prophylaxis conducted as part of an ongoing campaign to control and eradicate the disease.

1.4.3 Heterogeneity in mixing

As discussed so far the models we have presented all make one key assumption; that of homogeneous mixing. We have assumed that every member of the population mixes with each other at the same rate, thus every individual has the same risk of acquiring infection regardless of which individuals are infected. For human populations our intuition tells us that this assumption must be false. A politician on the campaign trail can interact with hundreds of people in a day, while a PhD student writing their thesis may have no social interaction at all.

Heterogeneity in contact structure can be introduced into disease modelling in a variety of ways. The main drivers of seasonal flu are school aged children. Through surveys it is possible to estimate how much mixing exists between school aged children and adults. The ODEs can then be extended to include this with the mixing between the age groups being encoded in a transmission matrix.

1.5 Network models

A natural way to introduce heterogeneity is to use network models. Networks can either be artificially created or assembled from data.

In disease terms a network is a collection of nodes that represent the base infectious unit. They can represent humans, animals, households, farms or any other division. The nodes are contacted to each other through edges which represent paths of possible transmission, generally through direct contact.

The network is encoded in an adjacency matrix, $\mathbf{A}$. The most basic network

has symmetric, binary edges, encoded simply in the adjacency matrix as

$$A_{ij} = A_{ji} = \begin{cases} 1, & \text{if a link between } i \text{ and } j \text{ exists} \\ 0, & \text{otherwise.} \end{cases} \tag{1.15}$$

An epidemic can then be simulated through the network by setting one (or more) nodes to be infectious. The nodes connected to the seed node can then become infected either deterministically or stochastically. The disease can then propagate through the edges of the network infecting susceptible nodes. In this way it is obvious that the structure of the network will greatly influence the spread of the disease.

1.5.1 Edge direction and weight

The simple binary network can be an appropriate representation of a data source. For some social media networks such as Facebook or LinkedIn each link between members is reciprocated, whilst for others such as Twitter one member can connect to another without that connection being returned. This is a directed network and is easily encoded in an adjacency matrix

$$A_{ij} = \begin{cases} 1, & \text{if a link from } j \text{ to } i \text{ exists} \\ 0, & \text{otherwise.} \end{cases} \tag{1.16}$$

We can note that the number of links in the directed graph is the sum over all rows and columns of $\mathbf{A}$ but in the directed case the number of links will be this sum divided by two.

Binary networks (with and without direction) give each edge that exists the same value and thus has the same level of importance in the network. In a disease context this may not be appropriate, we may want to adjust the strength of an edge to account for the relative risk of transmission along different links. These edges are said to be weighted and can also be encoded into an adjacency matrix

$$A_{ij} = w_{ij} \tag{1.17}$$

where w_{ij} is the weight (which may be zero) of a link from j to i.

1.5.2 Dynamic networks

A network does not have to be static. Edges can come and go (or increase and decrease in weight) as connections are made and broken. In both the modelling and analysis of livestock networks it is important to consider what time frame most represents the data [Vernon and Keeling, 2009]. The network could be constructed daily but then any models may suffer from random effects, if on the other hand the network is constructed yearly then important seasonal patterns may be missed.

1.5.3 Network analysis

Not only can we simulate the spread of a disease on a network but we can also study the structure of the network itself. This analysis can allow us to identify nodes and edges that may by targeted for removal or vaccination [Christley et al., 2005]. The statistics that we compute from networks may also be compared across different networks.

Reviews of network analysis techniques as applied to livestock disease have been performed by Dubé et al. [2009] and Martínez-López et al. [2009] while a comprehensive review on networks and infectious disease epidemiology in general was written by Danon et al. [2011]. A general introduction to networks from a more mathematical perspective is the book by Newman [Newman, 2010].

In chapter 2 we will discuss a variety of different network statistics and compute them for the cattle movement network in Turkey. The robustness of these statistics and ability of disease modelling to accurately predict epidemic size and spread when limited data is available is examined in chapter 3. A novel analysis of both networks will be undertaken in chapter 4 where we introduce the notion of a community in a network and perform community detection on the networks.

1.6 Modelling livestock disease

So far we have reviewed infectious disease modelling in a general sense. Much of what has been described has been motivated with its application to human disease. However the basic tenets of epidemiological theory remain true. First a decision as to whether the farm or the animal will be the basic epidemiological unit must be made. This decision will influence the choice of which classes should be considered and how heterogeneities in the contact structure should be accounted for.

1.6.1 FMD in 2001

Prior to the 2001 UK outbreak of FMD livestock models were not widely studied. This changed dramatically during the outbreak where modellers played a key in the control of the epidemics through advise to policy-makers.

Three main modelling groups put forward models. The Imperial model [Ferguson et al., 2001a,b] is deterministic, solving a set of ordinary differential equations. Interspread [Morris et al., 2001] is a highly detailed stochastic simulation model incorporating 54 parameters. The Keeling model [Keeling et al., 2001] is also stochastic but has a simpler structure than Interspread. A detailed discussion of the three models can be found in Kao [2002] and Keeling [2005a].

Since the 2001 outbreak other countries have developed models to investigate hypothetical outbreaks of FMD. NAADSM [Schoenbaum and Disney, 2003] and DADS [Bates et al., 2003] where both developed for the US, with DADS also recently being adopted for Denmark [Boklund et al., 2013]. Ausspread [Garner and Beckett, 2005] was developed for Australia while Interspread was refined as Interspread Plus [Stevenson et al., 2013] and is commercially available. Comparisons between various combinations of these models have been carried out by Dubé et al. [2007] and Halasa et al. [2014].

We now discuss in more detail the Keeling model which has the advantage over the Ferguson model of being stochastic, thus providing a range of possible epidemic trajectories, but has a more transparent structure than Interspread.

1.6.2 The Keeling model

The Keeling model treats the farm as the basic epidemiological unit and models the spread of infection between farms based on the spatial locations and livestock demographics of the farms involved. Each farm i has an associated susceptibility S_i and infectiousness I_i which are determined by the number of cattle n_i^{cattle} and number of sheep n_i^{sheep} and the relative susceptibility and infectiousness of that type of livestock

$$S_i = s^{\text{cattle}} n_i^{\text{cattle}} + s^{\text{sheep}} n_i^{\text{sheep}} \tag{1.18}$$

$$I_i = i^{\text{cattle}} n_i^{\text{cattle}} + i^{\text{sheep}} n_i^{\text{sheep}}. \tag{1.19}$$

The transmission process is stochastic such that on any day the probability of a farm i becoming infected is determined by the equation

$$P_i = 1 - \exp\left[-S_i \sum_{j \in \text{infectious}} I_j K(d_{ij})\right] \tag{1.20}$$

where $K(d_{ij})$ is the spatial kernel which defines how a farms infectiousness decreases with distance from a susceptible farm. The sum is over all infectious farms, creating an infectious pressure on the susceptible farm rather than explicitly having one farm infect another. The kernel subsumes all routes of transmission allowing for rapid parametrization of the model and calculates Euclidean distance between farms.

This work has had many extensions. Variations of the kernel using quickest route and shortest distance between farms were tested by Savill et al. [2006], though no benefit to using the original Euclidean distance was found. The the form of S_i and I_i in original Keeling model assumed that a farms susceptibility and infectiousness increased linearly with farm size. While larger farms will have potentially more infectious contacts than smaller farms this original assumption would make a farm with 100 animals 100 times more infectious than a farm with 1 animal. Tildesley et al. [2008] added and fitted power-laws to equation 1.19 to

$$S_i = s^{\text{cattle}}[n_i^{\text{cattle}}]^{p_s} + s^{\text{sheep}}[n_i^{\text{sheep}}]^{p_c} \tag{1.21}$$

$$I_i = i^{\text{cattle}}[n_i^{\text{cattle}}]^{q_s} + i^{\text{sheep}}[n_i^{\text{sheep}}]^{q_c}. \tag{1.22}$$

where $p_s, p_c, q_s, q_c < 1$ decreasing the effect of large farms sizes on the rates of infection.

In a formal parametrization of the model under a Bayesian framework Deardon et al. [2010] made use of the extension by Tildesley et al. [2008] as well as reshaping the kernel with a distance cut-off but introducing a 'spark' term to account for long range infection.

The original work of Keeling et al. [2001] was used to great effect during the 2001 outbreak since then the model and its various extensions have primarily used to experiment with control strategies and to assess the strategies implemented during the outbreak.

During the 2001 outbreak vaccination was not considered a viable control option. Since then many studies have been carried to determine the optimal control strategy. Keeling et al. [2003] considered both prophylactic and reactive vaccination while Tildesley et al. [2006] sought the optimal radius around a farm for reactive vaccination. The effect of pre-emptive culling during the outbreak was examined by

Tildesley et al. [2009]

Of course FMD is not the only livestock disease which has come to the attention of modellers. BSE [Kao et al., 2002], scrapie [Kao et al., 2007] and E-coli [Zhang et al., 2010] have been investigated but of particular interest is bovine tuberculosis (BTB). BTB is endemic in the UK and has drawn a lot of focus recently with regard to the controversial implementation of a badger cull by the UK government. BTB differs drastically from FMD in that it has a very long latent period, leading to significantly different dynamics.

The model by Brooks-Pollock and co-workers [Brooks-Pollock et al., 2014] has differed from the FMD models and the approach of Green et al. [2008] in that the animal rather than the farm has been used as the basic epidemiological unit. Transmission occurs between cattle on a farm and from local spread. Cattle move between farms thus infecting new farms if the cow in question is in the latent or infectious compartment. The CTS is once again used to determine which cattle should be moved.

We will follow the more traditional approaches of the previous FMD models when considering epidemics of livestock disease in the UK (chapters 3 and 4) and set the farm as our basic epidemiological unit. However, when we turn our attention to endemic FMD in Turkey in chapter 5 we will focus on the within herd dynamics of FMD outbreaks and model individual cattle.

1.6.3 Cattle movements

The 2001 FMD outbreak and the Keeling model [Keeling et al., 2001] have given rise to a wealth of literature and fascinating insights on how to model the spread of FMD. However, it must be stressed that the Keeling model captures the 2001 outbreak following the enforcement of a movement ban, thus spread is generally localised. An in depth investigation has since revealed that FMD was imported by a farmer in Norwich who moved pigs to market, from there it spread across the UK. The outbreak was detected sometime after this and thus the initial spread of the infection was due to movement of livestock [National Audit Office, 2002].

Owing to this fact more recent models tend to incorporate livestock movements explicitly into the model, with some, though not all also including a spatial kernel as another route of transmission, [Kao et al., 2006] . Generally the farm is still kept as the base epidemiological unit with farms becoming infected through both the cattle movement network and through local spread, though the question of how to best deal with movements cancelled in the event of a movement ban remains an open one.

The availability of the cattle tracing system (CTS) in the UK has afforded modellers accurate data to build these network based models. The CTS came into being in 1998 following the outbreaks of bovine spongiform encephalopathy (BSE) and was mandated by EU law which required all member states to have a computerised tracing system. All commercial movements of every cow, including movements through a market are included. Since the 2001 outbreak the quality of this data has increased as recording became compulsory for all cattle [Green and Kao, 2007].

Modellers can make requests to Defra for extracts of the CTS data. This can come as individual movements charting the full history of each cow, or as a table of batched movements giving the off-holding, on-holding, date of movement and the number of cattle moved in the batch.

This CTS data has allowed for a detailed analysis of the cattle trade network in the UK, [Green et al., 2006; Kao et al., 2006; Kiss et al., 2006; Tildesley et al., 2011], which will be discussed in chapter 2. As recording of cattle movements is a European directive studies have also been carried out in Sweden [Nöremark et al., 2009], Italy [Bajardi et al., 2012], and France [Dutta et al., 2014; Rautureau et al., 2012]. We will also perform an analysis on the Turkish cattle movement network.

While the availability of CTS is a great asset to those modelling livestock disease in the UK, not all countries have the resources to maintain such a detailed system. In chapter 3 we assess the viability of making epidemic predictions if only a partial sample of such data were available.

1.7 Aims of the thesis

The purpose of this thesis is to analyse the cattle movement data from Turkey in the context of FMD with the goal of aiding the Turkish authorities in the control and eventual eradication of the disease. Each chapter contributes towards this goal. First we introduce the data and discuss its quality and the processes of cleaning the data. By applying network analysis we can find identify if there has been inherent changes in the trade of cattle in Turkey which could be drivers that may reduce or enhance the spread of FMD.

Chapter 3 serves a dual purpose. By focussing on the UK, a country that we know to have near complete cattle trade data, we can explore how close to completeness a data set must be for us to be confident that any analysis of that network will be robust. Secondly in exploring various sampling techniques we aim to give policy makers a tool for guiding data collection should the collection of complete data be too costly.

We further enhance our knowledge of both the British and Turkish cattle movement networks by applying community detection methods to these networks. In doing this we seek to gain deeper insights into the network structure than those afforded us by traditional network analysis. We also aim to explore community detection as a a technique to predict which groups of farms are likely to become infected in an epidemic adding another tool to the policy makers arsenal alongside locally, regionally, or nationally targeted control measures.

In constructing a within herd model of FMD we build on the modelling literature, adapting existing compartmental models to suit the epidemiology of the disease taking into account the cattle farming practices in Turkey. By focusing within herd we seek to gain a handle on the key parameters that influence the spread of FMD in an endemic setting.

Chapter 2

The data

> "It is a capital mistake to theorise before one has the data."
>
> Sir Arthur Conan Doyle, A Study in Scarlett [1887]

After consultation with the Turkish Veterinary authorities facilitated by the European Commission for foot-and-mouth disease (EuFMD) we were given access to data from the TurkVet database. In this chapter we will explain how the data were processed and cleaned. We will also analyse these data statistically, graphically and using networks.

The most resolved cattle farming unit in Turkey is the holding of which there are over 2,990,582 and the birth, movement and death data are recorded at this resolution. For comparison the UK has approximately 70,000 cattle farms. The number of animals on these holdings can range from fewer than 5 to over 500. As many of these holdings are small the basic epidemiological unit for recording disease spread in Turkey is not the holding but the epiunit. An epiunit is a village or a neighbourhood comprised of several holdings. For simplicity when comparing with the UK and other countries we will refer to an epiunit as a farm. Histograms of the number of holdings per farm are shown in figure 2.1 as is the number of births per holding in 2010.

2.1 Raw data

First we created a database in PostgreSQL to handle the large quantities of movement data available from the TurkVet database. The basis behind an SQL based database is to create tables of data and to link and query these tables appropriately. The bulk of the data came in the form of a .dmp file from TurkVet. This file was extracted into 1491 .xls files, each of which had 65,535 lines of movement data save

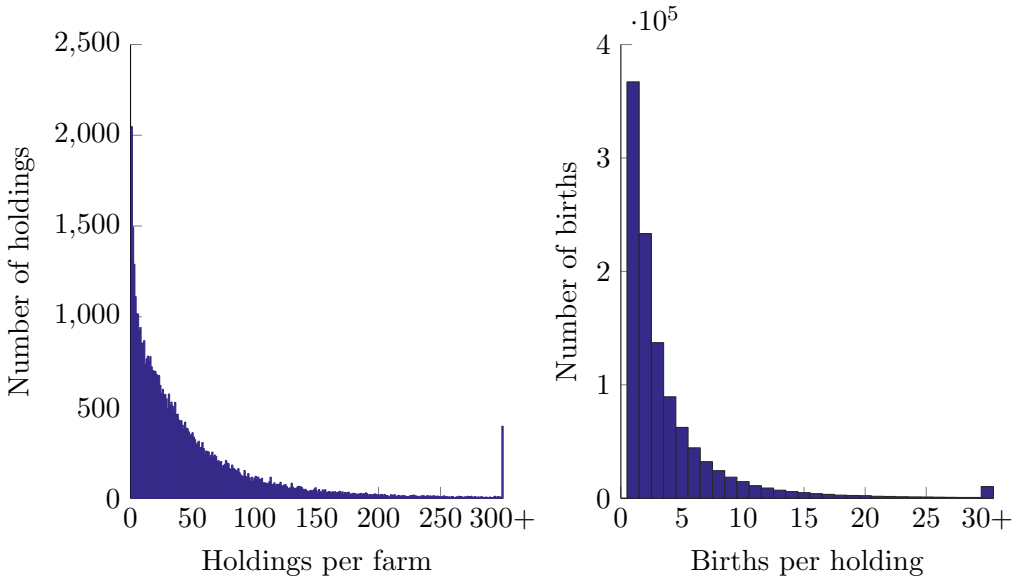

Figure 2.1: Histograms of the number of holding per farm (left) and number of births per holding in 2010.

the last file that had 30,613 giving a total of 97,677,566 lines of data. The column headings for each of these are given in table 2.1.

#	Column Heading	Description
1	Animal ID	Cattle ear-tag number
2	From Province	The province of origin
3	From District	The district of origin
4	From Neighbourhood	The farm of origin
5	To Province	Mistake here, this is actually the holding of origin
6	To District	The district of destination
7	To Neighbourhood	The farm of destination
8	To Holding	The holding of destination
9	Departure date	The date of departure
10	Arrival date	The date of arrival

Table 2.1: Data fields in the event database

The cattle ear-tag is unique to each cow. These consist of a two character country code followed by twelve digits. For the case of Turkey the first two digits indicate the province of the cow's birth. The country code for Turkey is TR.

The TurkVet database records cattle movement and cattle birth. Births are indicated by the origin and destination holding being the same. Separating the

Figure 2.2: Cow TR420001738341 taken at a cattle market in Konya by the author in January 2014. The Turkvet records made available to us show that this cow was born on the 25th January 2011.

births from the movements gives us 49,870,920 movements and 47,806,646 births.

The departure date and arrival date are the same for 45,443,954 movements ≈91% of all movements. Of the remaining 4,426,966 entries, 255,061 arrived the day after departure and another 1,758,918 in the following week. However 7,684 are recorded as arriving before departure. It would appear that most of the remainder of the discrepancies arise due to clerical error. For instance 15 cattle supposedly arrived 21,195 days (sixty years) before departure but on querying this we find that the departure date was given as 17/12/2066 with arrival date as 17/12/2006. For other cases when it is less obvious which is the correct date to consider we can look at the other movements by the cattle, choosing the date that ensure causality. Batches of cattle moved on the same day between the same holdings seem to have the same errors, finding and fixing the error for one cow should enable fixing for all.

2.1.1 Death data

The next substantial piece of data we have available are cattle death records. This came in a similar format to the event data and when extracted from the .dmp file gave 159 .xls files with approximately 65,535 lines in each, giving 10,394,299 death records in total. The headings for each column are given in table 2.2.

The fourth column indicates the cause of death and these are given in table 2.3. Of these causes #8 accounts for almost 7 million of the 10 million deaths and refers to cattle slaughtered en masse at a slaughter house. The next most common with 1.6 million records was #3 which is unknown. Animals specifically killed for the Kurban festival #10 account for almost 1 million more records.

#	Column Heading	Description
1	Eartag	Cattle ear-tag number
2	Death Type	Cause of death
3	Holding	Holding where the animal died
4	Death date	Recorded date of death

Table 2.2: Data fields in the death database

#	Turkish	Rough Translation
1	'ANAFLAKTIK SOK'	Anaphylactic shock
2	'AYIRIP OLDURME'	Take the killing
3	'BILINMEYEN'	Unknown
4	'DOGAL OLUM'	Natural death
5	'HASTALIK SONUCU'	As a result of disease
6	'KANSER'	Cancer
7	'KAZARA OLUM'	Accidental death
8	'KESIMHANEDE IMHA'	Disposal slaughterhouse
9	'KOMBINADA KESIM'	Combined cutting
10	'KURBANLIK KESIM'	Kurban sacrifice
11	'TASIMA SIRASINDA OLUM'	Death during transport
12	'ITLAF'	Cull
13	'SAHSI KESIM'	Individual cutting
14	'SAHSI IMHA'	Individual disposal

Table 2.3: Causes of death with rough translation

2.2 Processing the data

To begin the process of data cleaning and data analysis the vast sums of data were imported into a custom made database run through PostgreSQL version 9.3.5. The large number of events (births, movements and deaths) made handling these data through conventional software such as MatLab impossible due to memory limitations. SQL based databases use tables as their main data storage containers. A table was created for each of the three event types. From these initial tables we could combine entries to create a separate table which would give summary information on each individual animal, such as ear-tag, date and place of birth, number of moves and date and place of death. Joining the event tables together gave 49,307,854 unique cattle ear-tags.

When creating the tables it is good practice to create a primary-key for that table which uniquely indexes every row in that table. Another table was created to hold the indexes for each of the events in the other tables along with the index for

# Births	# IDs	Proportion
0	1588845	0.032
1	47641339	0.966
2	69926	0.001
3	6637	0.000
4	474	0.000
5	250	0.000
6	291	0.000
7	87	0.000
8	2	0.000
9	3	0.000

Table 2.4: Number of recorded births per ear-tag.

the animal involved in that event. This table had 108,071,865 entries one for each event and was vital for the quick referencing of events across the different tables.

Now that a 'cattle' table and an 'events' table existed it was possible to begin identifying events that needed to be cleaned. A routine was put in place in MatLab that would query each individual animal from the database and return all events that animal was involved in. The events would then be checked to make sure that they happened chronologically between holdings, and that the chain of holdings was consistent. Cattle which had errors such as the event chain being broken, multiple births or deaths were identified, to be dealt with later.

2.2.1 Births

In theory cattle ear-tags are unique so there should never be more than one recorded birth per ear-tag. As seen in table 2.4 the majority of cattle in the database, 96.6%, have a birth record. Of the remainder, 3.2% have no birth record, 0.1% have two and 7,519 cattle (less than 0.001%) have multiple birth records.

The temporal distribution of the birth records can now be explored. This can be done at different resolutions, first the number of births recorded yearly (figure2.3). The number of records has grown steadily since 2000 with a figure of 4-5 million births a year being recorded consistently since 2006. The pre-2000 data show some records which clearly contain errors, with two cattle showing as being born in 1900. The latest birth date was 8th March 2013. We now focus on the birth records from 2006 to 2012.

Cattle are bred seasonally so we would expect to see seasonal variation in the number of birth records. In figure 2.4 the birth records from 2006 to 2012 are

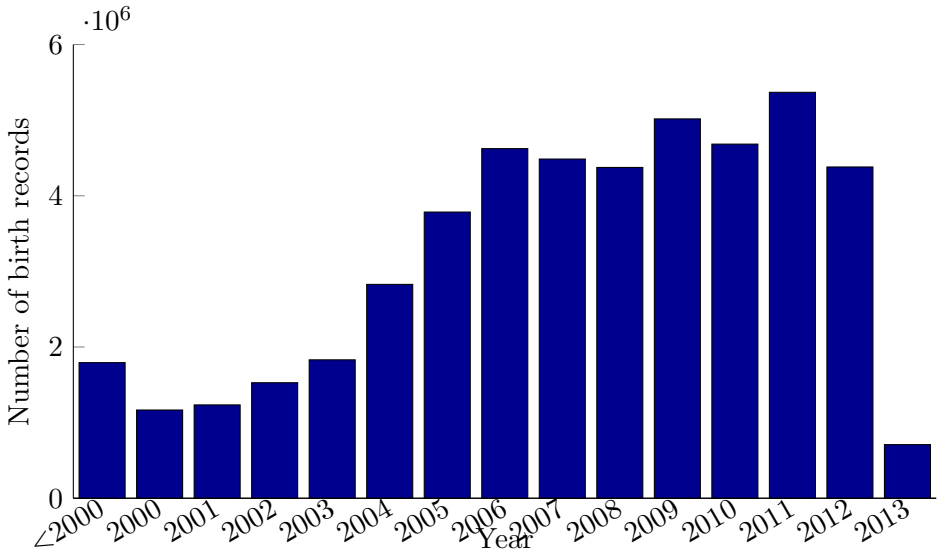

Figure 2.3: Bar chart showing the number of births recorded per year. The leftmost bar accumulates all records before the year 2000. The 2013 records date until 8$^{\text{th}}$ March.

broken down monthly. The greatest number of births are recorded in January for the years 2006 to 2010. February and March also have high numbers of records apart from in 2009 when the peak month was May. The number of birth records in every month is seen in the stacked bar chart (figure 2.5). This shows that the accumulative number of births is higher at the start of the year with a minimum around August.

The final resolution we can probe is the number of birth records per day. Figure 2.6 shows the number of birth records per day from 2007 to 2012. The most striking feature of these bar charts is the spike that appears on the 1$^{\text{st}}$ January every year and the smaller spikes that appear on the first day of every other month. These spikes show a clear reporting bias presumably arising from veterinarians or farmers not recording the exact birth date correctly.

The extent of this bias can be seen in figure 2.7. Here we plot a histogram of the proportion of births recorded on the first of the month in each farm. For comparison with a random day we also make the same plot but for births recorded on the 25th of the month.

If all days were equal one would expect the proportion of births on any given day to be distributed around $\frac{1}{30} = 0.333$ with farms with few births populating the tails of this distribution. We can see that this is the case for births on the 25th

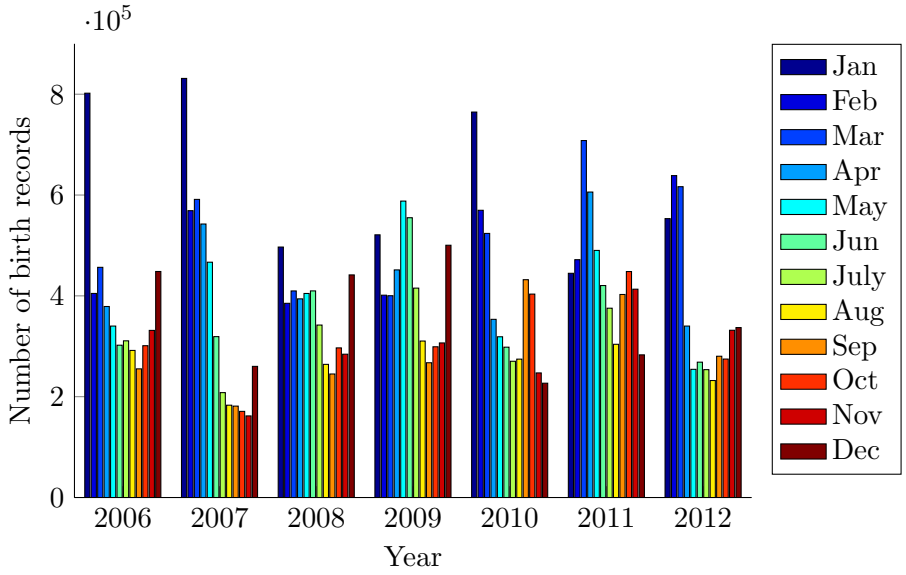

Figure 2.4: Bar chart showing the number of births recorded per year divided into months for the years 2006-2012.

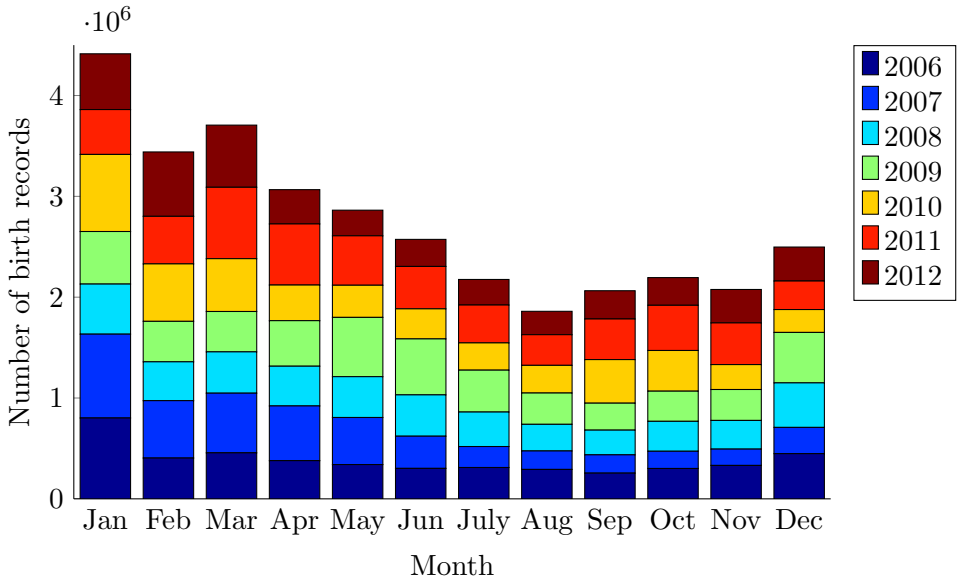

Figure 2.5: Bar chart showing the number of births recorded per month for the years 2006-2012.

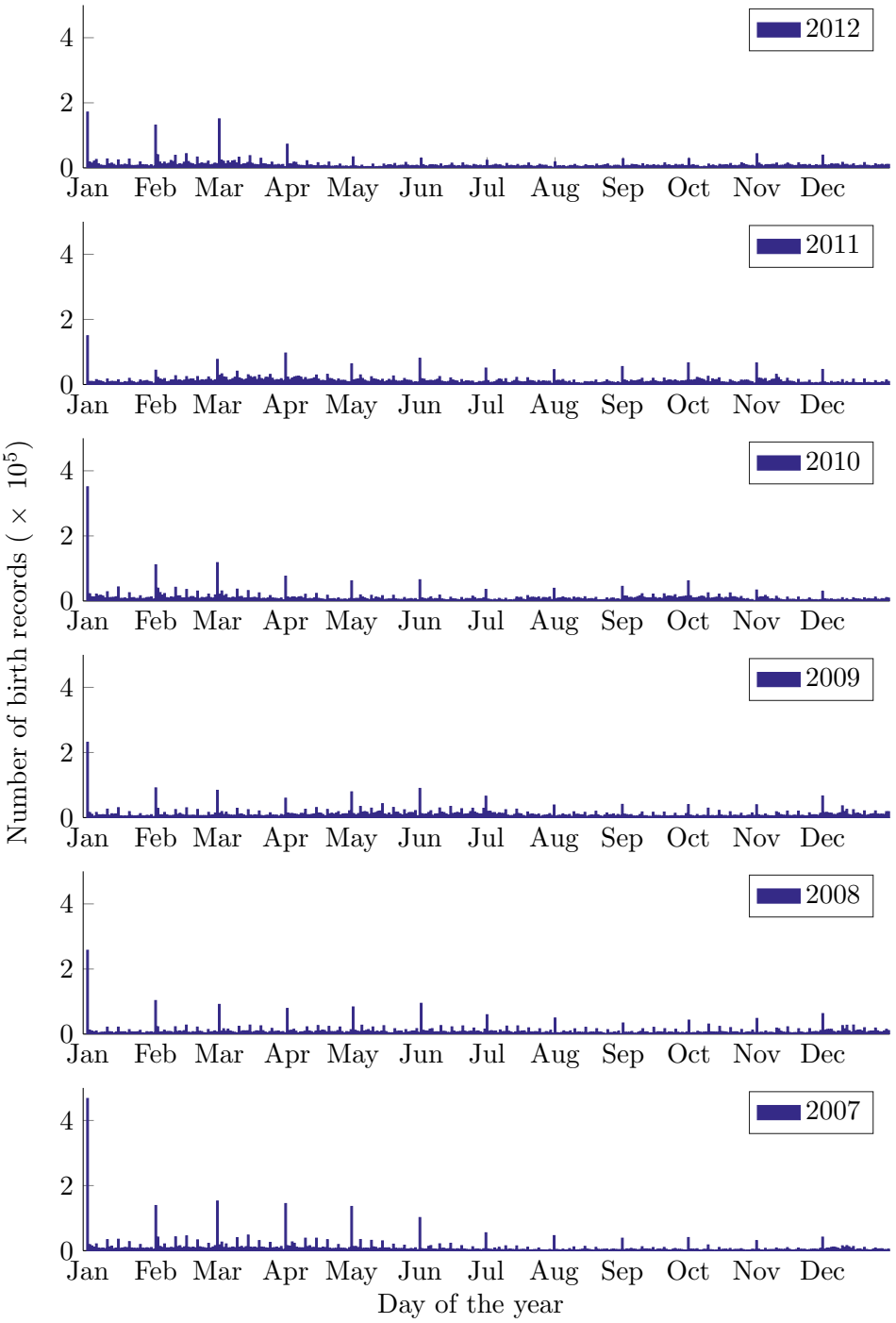

Figure 2.6: Bar chart showing the number of births recorded per day for the years 2006-2012.

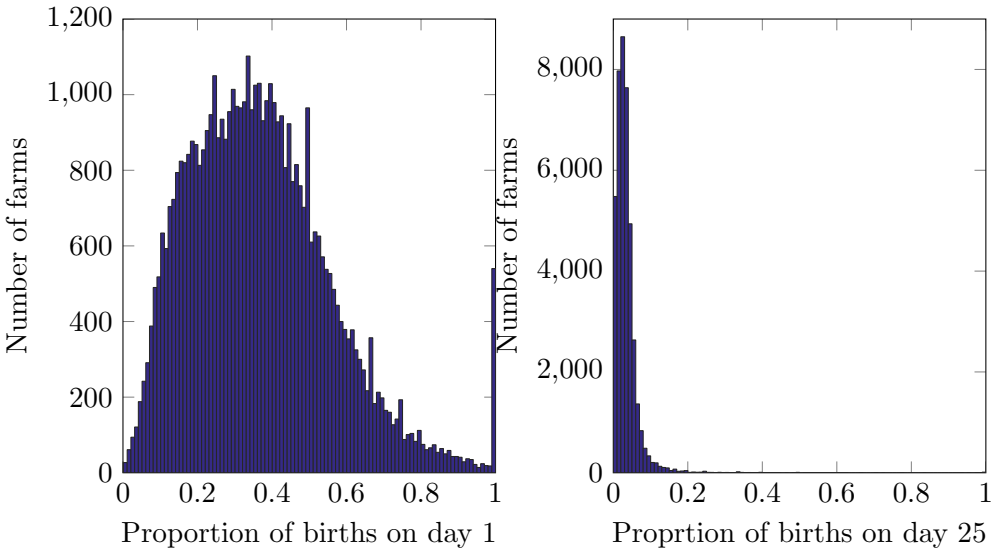

Figure 2.7: Histograms of the number of births per farm recorded on the 1st and 25th of the month.

with the distribution tightly centred around 0.2, while for the 1st the distribution is wide with the majority of farms recording 40% or higher of their births on the 1st. Indeed many farms record all of their births on the 1st.

2.2.2 Death records

Death records do not appear to have been kept as well as the birth records. The number of deaths per year are plotted in figure 2.8 and are much less than the number of births, indicating an increase in the yearly cattle population. Indeed comparing figure 2.3 and figure 2.8 we would expect the cattle population to have increased by at least 2 million cattle a year since 2007, yielding a total increase of over 12 million cattle by the end of 2012. However, the official statistics state the cattle population has actually dropped slightly from 14 million in 2007 to 12 million in 2012[i].

2.2.3 Movement records

The movement data can be analysed in the same way as the birth records. However, we can choose to focus on either individual cattle moved between farms or on batches of cattle moved. The distribution of batch size from 2007 to 2012 is plotted in figure

[i] `http://www.turkstat.gov.tr`

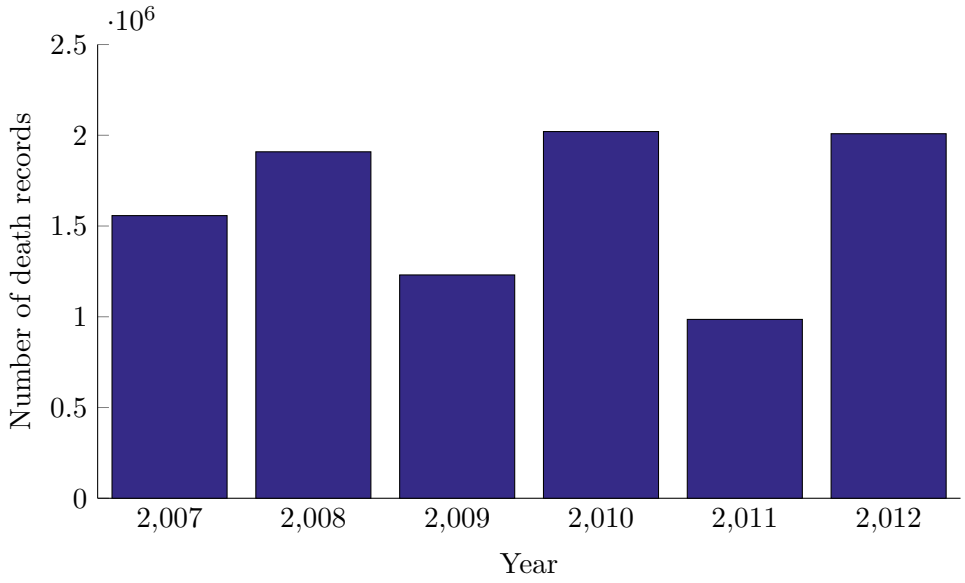

Figure 2.8: Number of annual death records from 2007 to 2012.

2.9. Over 60% of the batches contain only one cow with another 35% of cattle being in batches of size two to ten.

The Kurban festival

The Kurban festival ('Kurban Bayramı') is an Islamic festival which honours Abraham's willingness to sacrifice his son Issac [ii]. Abraham's gesture is commemorated through the sacrifice of animals. The meat from these animals is often distributed amongst the poor.

The Kurban festival moves each year but in the years we are considering it falls between October and December. Large numbers of cattle are moved to urban areas in the weeks leading to the festival in preparation for slaughter. We would expect to see peaks in the movement data corresponding to movement for this slaughter.

Seasonality in the movement records

With the majority of cattle being moved in such small batches we will focus on the number of batches rather than the number of cattle. At the monthly level we see that the number of movements is seasonal with one peak in April and May and a second peak in October and November, figure 2.10. Though the Spring peak is

[ii] https://en.wikipedia.org/wiki/Eid_al-Adha

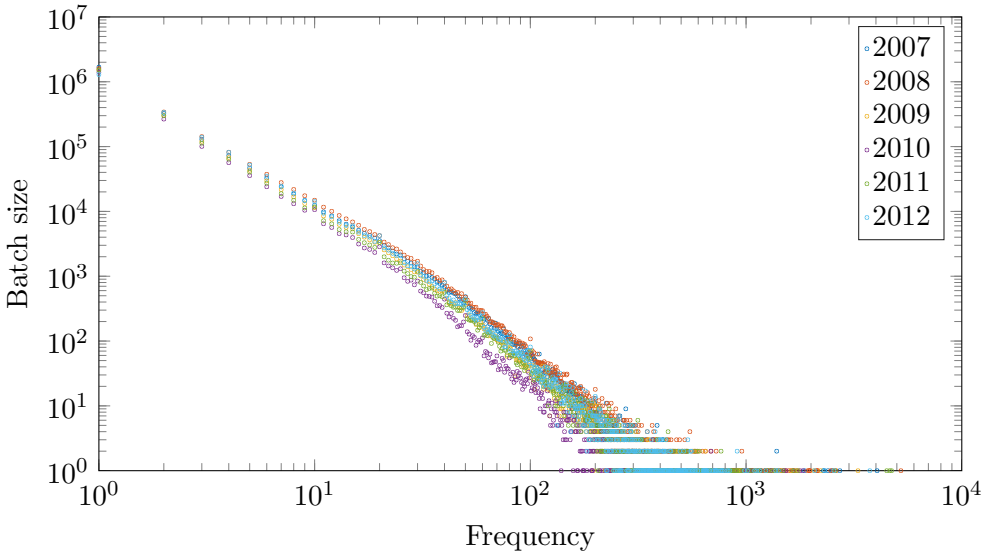

Figure 2.9: Distribution of batch sizes from 2007 to 2012.

generally higher than the Autumn peak for the number of batches, more cattle are moved in the Autumn, this is shown for the 2012 movements in figure 2.11 and for all years in appendix A (figure A.2).

By plotting the number of movements per day (figure 2.12) we see that the reporting bias evident in the birth data does not present itself in the movement data. Weekly variation can be identified with a dip at the weekends.

2.3 Cleaning the data

Using the database we were able to check each cow's movement history to see if any errors were present. As the model will be built on daily time-steps we wanted to ensure that no two events occurred on the same day. An example of how we altered the movement records for one cow is explained below with the original and edited history given in table 2.5.

First this cow has two recorded birth entries. Both these births take place at the same holding but at different dates. We choose to keep the second entry.

Records are reordered such that the from-holding on one record matches the to-holding on the previous record. In this example, as the to-holding in 3 is the same as the from-holding in 4 and both records occur on the same day we can merge the two movements together. In most cases the record would just have been edited. However we note further that the from-holding in 3 is the same as the to-holding in 4. Merging these creates a loop, so we completely remove both the records.

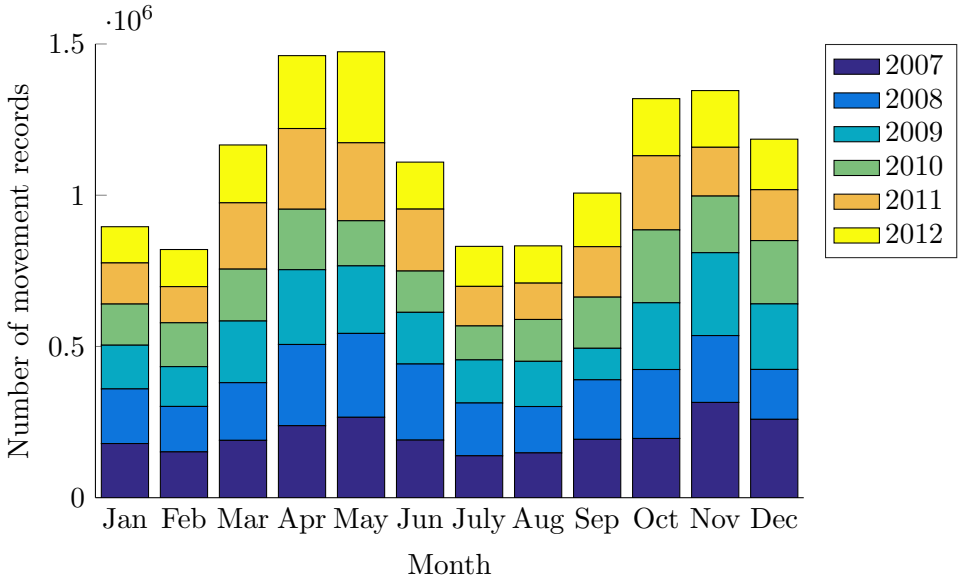

Figure 2.10: Bar chart showing the number of movements recorded per month for the years 2076-2012.

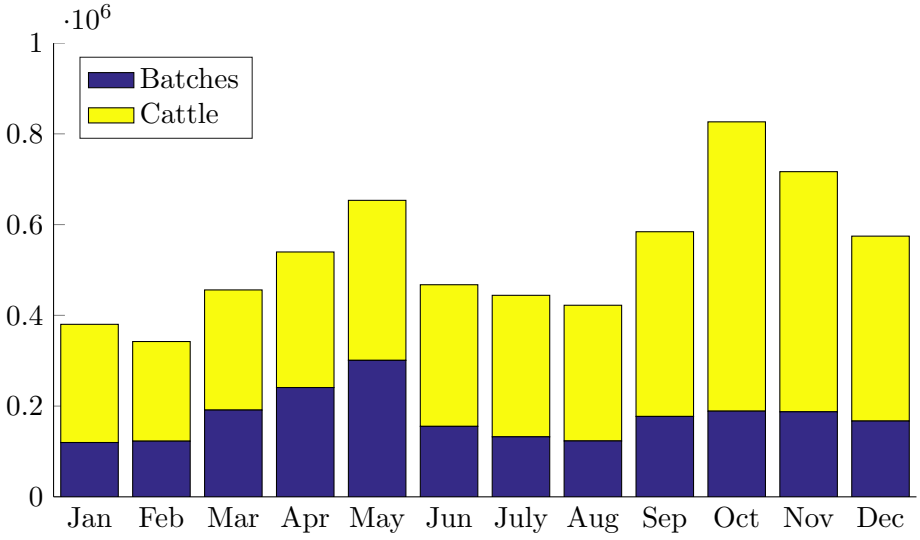

Figure 2.11: Monthly number of cattle moved and number of batches 2012.

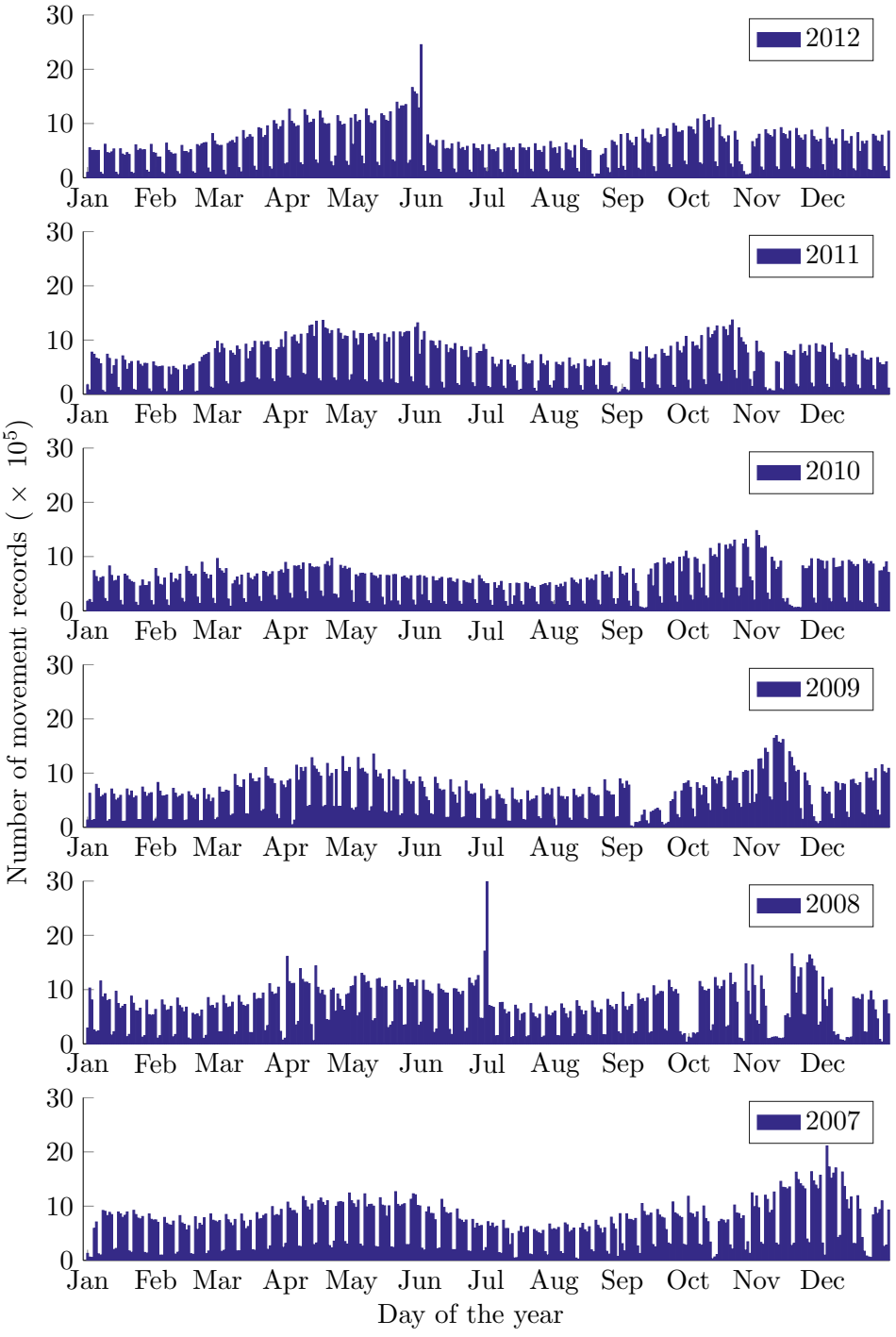

Figure 2.12: Bar chart showing the number of movements recorded per day for the years 2007-2012.

Original			
#	from	to	date
1	birth	1842444	01/02/2006
2	birth	1842444	14/05/2007
3	1842444	1813237	27/09/2007
4	1813237	1842444	27/09/2007
5	1842444	1813237	14/05/2008
6	1813237	1829052	31/05/2008
7	1829052	death	31/05/2008

Altered			
#	from	to	date
2	birth	1842444	14/05/2007
5	1842444	1813237	14/05/2008
6	1813237	1829052	31/05/2008
7	1829052	death	01/06/2008

Table 2.5: History of cow TR460000323069

Now that records 3 and 4 are deleted record 5 becomes the next record in the sequence, the from-holding in 5 matching with the to-holding in 2. Record 6 then follows record 5. Record 7 is a death record and this has the same date as record 6. We could merge these records but that would change the farm where the death was recorded. Instead we simply add a day to the date record.

This process was carried out systematically for all cattle which had a non-causal event history.

2.4 Visualising the data

In order to visualise the data on a map we first had to find out where every farm was. A .xls file (*file1*) was supplied with the movement and death data. This file contained a row for every farm that appeared in the database; 55,193 in total. These farms were identified with an ID number ranging from 1-55,193. The name of the Neighbourhood of the farm was supplied as were the district and province names along with district and province ID numbers.

As well as *file1* three other .xls files were also supplied. Two files *file2* and *file3* contained coordinates for the neighbourhoods; *file2* contained the same information as *file1* for 55,088 neighbourhoods but also included latitude and longitude for 27,730 of these. Latitude and longitude for 39,347 neighbourhoods was contained in *file3* but in this file there were no IDs for the neighbourhoods and districts (only names). A final file (*file4*) contained large ruminant population numbers for 47,801 neighbourhoods from 2010 with longitude and latitude for 26,858 of them.

Neighbourhoods across the different data sets were matched by ID where available and string matching was then used to match the remaining entries. For those neighbourhoods that did not have a specified longitude and latitude a protocol

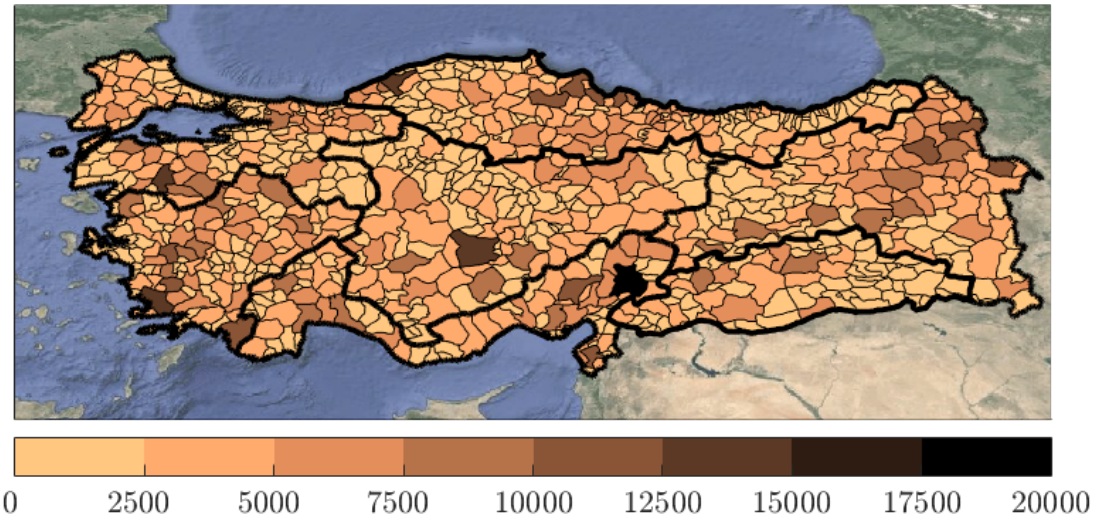

Figure 2.13: District level map showing the number of holdings in each district of Turkey. Provinces are outlined in bold.

was written in Python to query Google Maps with neighbourhood names. Latitude and longitude were extracted from these queries.

Across all files with a provincial ID that ID corresponded to a standard licence plate IDs for the 81 provinces. District identification proved slightly more problematic. According to Wikipedia [iii] there are 957 districts of Turkey, from the files we listed there are 980 districts and the shape-file we were supplied had 928 districts.

Each province in Turkey has a central district in charge of administration. Discrepancies between the data sources occur depending on if this central district has been subdivided or not. By referring to the Wikipedia list of districts and those given by the shape-file it was possible to assign each neighbourhood to a district and cross check this using the latitude and longitude coordinates for that neighbourhood.

An example of the maps we can create is given in figure 2.13. This figure shows the number of holdings in each of the 928 districts of Turkey. The background satellite image is taken from Google Maps.

2.4.1 Movement maps

The map of the holdings (figure 2.13) gives us some idea of the spatial spread of cattle around Turkey, though as holding size and farming practices will vary throughout the country this map may not give the best indication of disease risk.

[iii] `https://en.wikipedia.org/wiki/List_of_districts_in_Turkey`

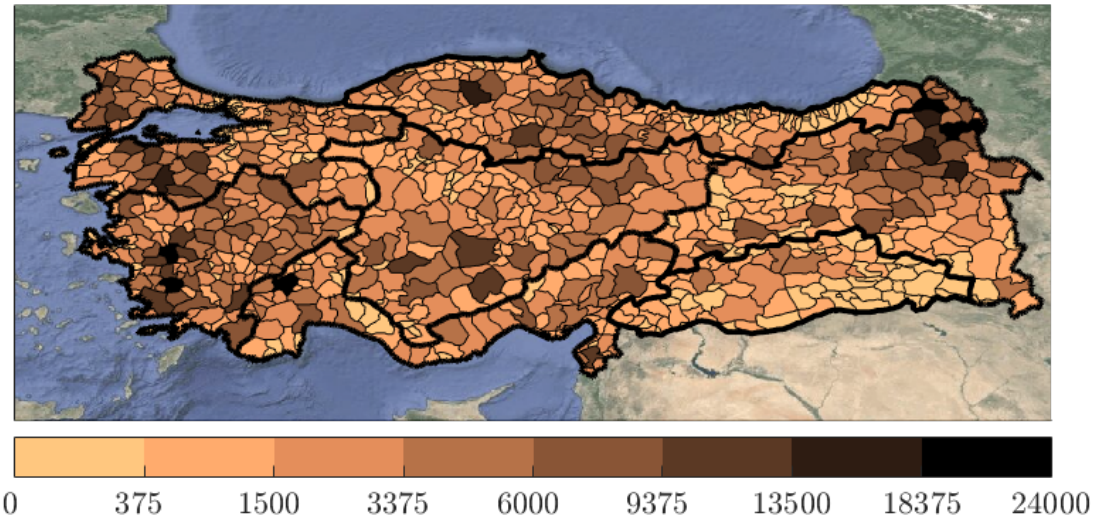

Figure 2.14: The number of movements from each district in Turkey for the year 2010. This includes the within district movements. Note that the colour scale is non-linear so as to show more variation between areas of high and low movement.

As we have identified in the introduction movement of animals is the main driver of the spread of FMD. Insight can therefore be garnered by plotting where animals move from and to. At a district level we can also examine which areas are most likely to trade locally or engage in long distance trades which can lead to nationwide outbreaks.

Figure 2.14 shows the number of movements originating in each district for the year 2010. There is a lot of activity along the West coast of the country and in the North-East. The South-Eastern Anatolia region near the Syrian and Iraqi border has little activity as do parts of the Mediterranean and Black Sea coasts. A similar map for the movements into each district highlights the same spatial patterns and is provided in appendix A (figure A.1). The same trends were also seen in the other years analysed.

Changing movement patterns

Movement patterns in Turkey appear to have shifted between 2007 and 2012 with a higher proportion of movements occurring within districts than was previously the case. Indeed in 2007 67% of movements were intra-provincial. For 2008 and 2009 this figure stayed relatively constant at 68% and 67% respectively. The shift occurred in 2010 when 70% of movements were intra-provincial and has risen to 74% in 2011 and 77% in 2012. The percentage of within district movements has increased even more dramatically from 32% in 2007 to almost 50% in 2010. As

	Intra-district	Intra-provincial	Total moves	Cattle moved
2007	32%	67%	2.5m	7.2m
2008	32%	68%	2.6m	8.1m
2009	34%	67%	2.3m	6.6m
2010	41%	70%	2.1m	5.9m
2011	44%	74%	2.4m	7.2m
2012	49%	77%	2.2m	7.9m

Table 2.6: Table summarising the numbers of movements and cattle moved for each year as well as the percentage of movements that were intra-district and intra-provincial.

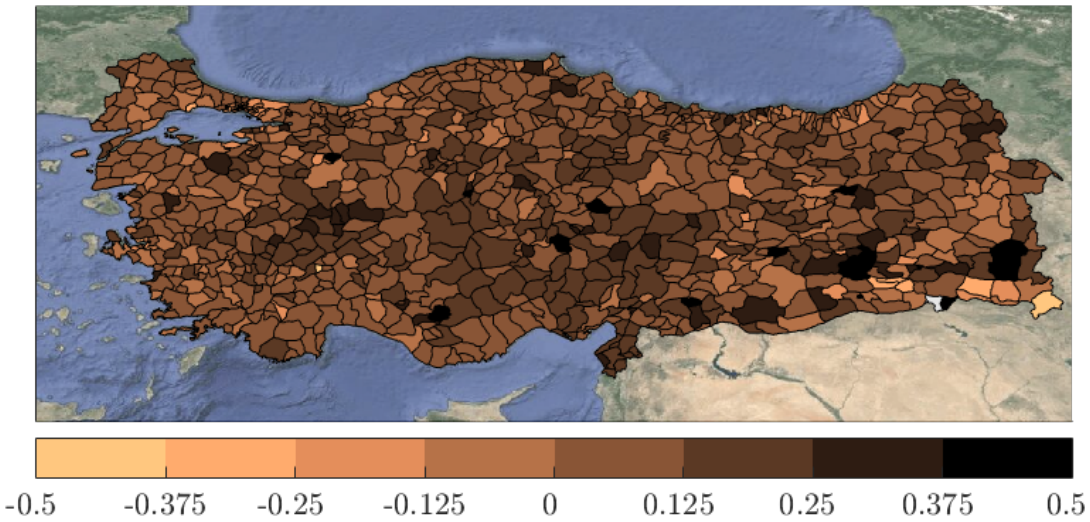

Figure 2.15: The difference between the proportion of intra-provincial movements for each district between 2007 and 2012. Darker colours indicate a high increase in the proportion of intra-provincial movements.

summarised in table 2.6 this increase in local level movement comes despite the number of movements remaining largely constant.

This increase in intra-provincial movements between 2007 and 2012 is shown in figure 2.15. Most districts undergo an increase in intra-provincial movements of between 5% and 15%, though some districts in the South East of Eastern Anatolia do report a decrease.

Borders and cities

Mapping the movement data allows us to identify hotspots in the country which receive many more cattle than they produce. One would expect such areas to be urban and represent the large cities.

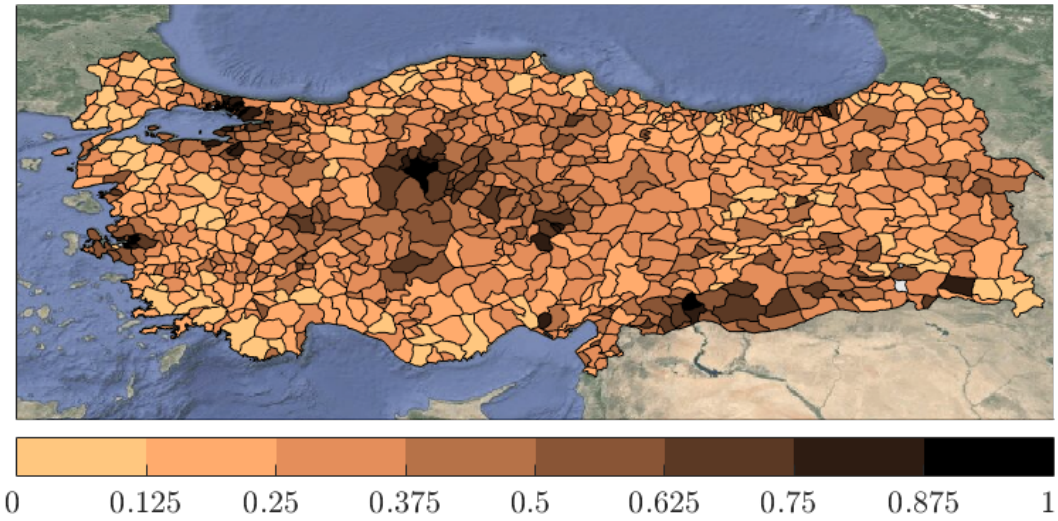

Figure 2.16: The proportion of long range movements received by district against the total number of movements destined for each district. Dark spots indicate Istanbul, Ankara in the centre of Central Anatolia, Izmir on the Aegean coast and the border with Syria and Iraq in the South.

We can map the proportion of long range movements that a district receives to highlight areas that mainly import cattle. Here we define a long range movement as a movement that did not originate in the district itself, its neighbouring districts, nor the districts that neighbour its neighbouring districts.

The map of long range movements is shown for 2010 in figure 2.16. The urban centres of Istanbul, Ankara and Izmir are easily identified. As well as highlighting the three largest cities in Turkey this map also identifies the border with Syria and Iraq as receiving a high proportion of long range movements. This region is sparsely populated. Indeed as identified through discussion with the Turkish Ministry of Food, Agriculture and Livestock this situation arises from a disparity between the price of cattle in Turkey and across the border.

As cattle were more expensive in Syria and Iraq, cattle would be moved to the border. These movements are legal and were thus recorded in the TurkVet database. The cattle would then be illegally moved across the border.

Similar maps of long range movements pick out the cities and borders, though the extent of movement to the border seems to have waned by 2012.

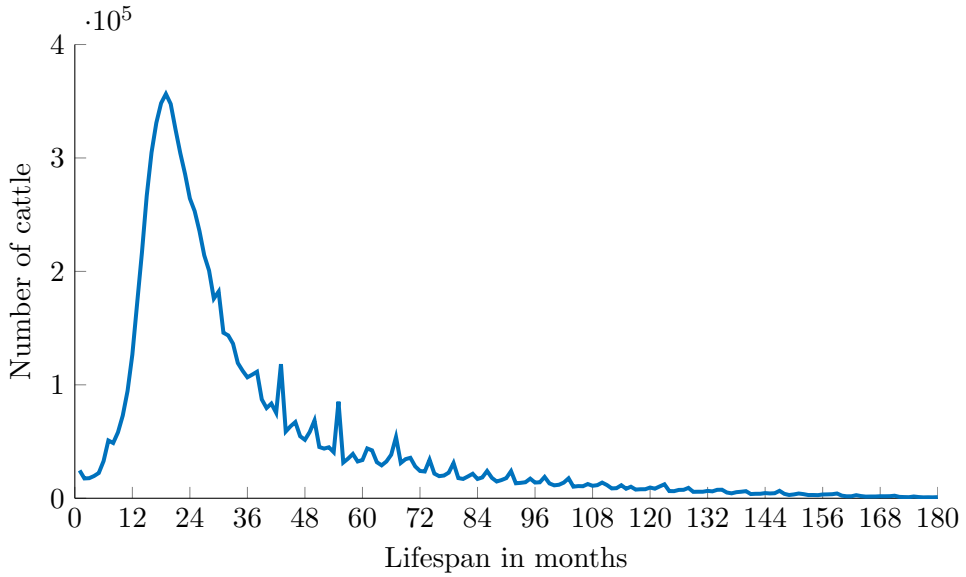

Figure 2.17: Distribution of the lifespan of cattle based on cattle with a recorded dates of birth and death.

2.5 Demography

A clear problem with the data is the under reporting of deaths. Using the cattle that had both a recorded birth and a recorded death we can built a distribution to describe the lifespans of cattle in Turkey, figure 2.17. There is a peak at 19 months and almost 24% of cattle died between 18 and 24 months. The age cattle live to drops off quickly after 20 months with very few cattle living beyond 15 years.

This distribution can be used to project how long a cow will live based on its ages and allows us to assign dates of death to all cows that have no death record. We do this by counting the number of cattle with a birth record but no death record born in every month from December 2012 counting back to January 2000, N_m.

Drawing from a multinomial distribution with N_m trials and the probabilities described by normalising the curve shown in figure 2.17 we can model how many cattle born in a certain month will die in every subsequent month. We can store this information in a matrix $\mathbf{E}$ where each element $E_{m,d}$ is the number of cattle born in month m that we estimate will die in d months time and $\sum_d E_{md} = N_m$.

We also count the number of cattle with both a birth and a death record and store this information in a second matrix $\mathbf{R}$ which is equivalent to $\mathbf{E}$ with the elements $R_{m,d}$ representing the number of recorded births in month m with a recorded death d months later. We then sum the recorded and estimate life span matrices to give a total life span matrix $\mathbf{T} = \mathbf{E} + \mathbf{R}$. A row of this matrix gives the

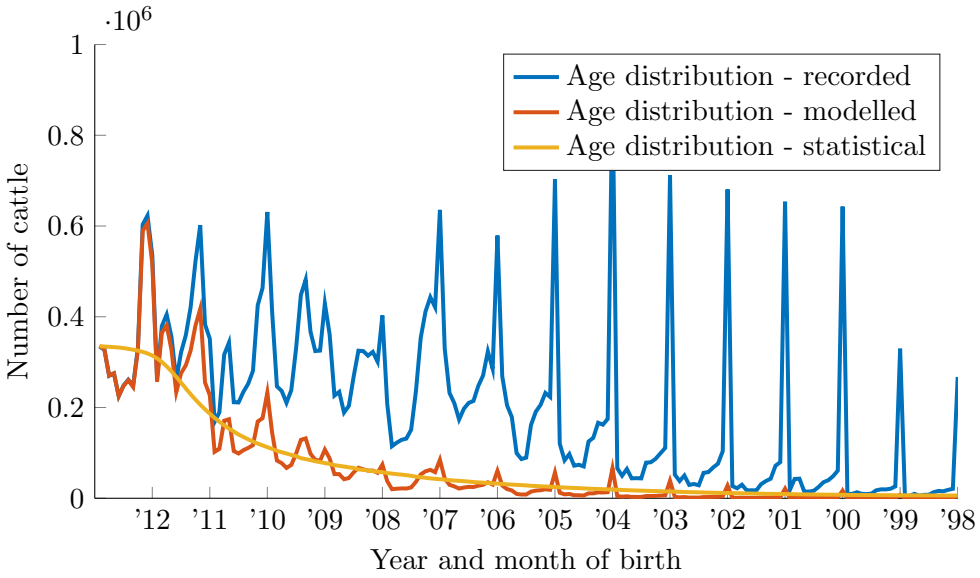

Figure 2.18: Age distribution of cattle in Turkey from the data, edited using the lifespan distribution combined with survival analysis, and a theoretical distribution that directly using the cumulative distribution of the lifespan distribution.

lifespan distribution for all cattle born in that month.

Finally, we can use $\mathbf{T}$ to obtain the age distribution for all alive cattle at any month by creating a new matrix $\mathbf{A}$ such that

$$A_{m,d} = \sum_j T_{m+d-1,j}. \tag{2.1}$$

The $m = 1$ row for this matrix gives our estimated age distribution for cattle alive in December 2012.

The results of this modelling of the age distribution are given in figure 2.18. This approach removes many older cattle and allows for the natural seasonality of the data to remain. We can now calculate the total number of cattle we expect to be alive at the end of December 2012, this comes to 13.0 million.

An alternative approach would have been to directly apply the cumulative lifespan distribution scaling by the number of cattle born in December 2012. This theoretical distribution has no capacity to take into account seasonality. This theoretical approach estimates the number of cattle alive in Turkey to be 12.8 million.

These population estimates will vary depending on which month is used as the present day. We can sum over the columns of $\mathbf{A}$ to give a population estimate for each month. In figure 2.19 we calculate the population estimates for each month from December 2012 back to January 2010. Our estimates are compared with the

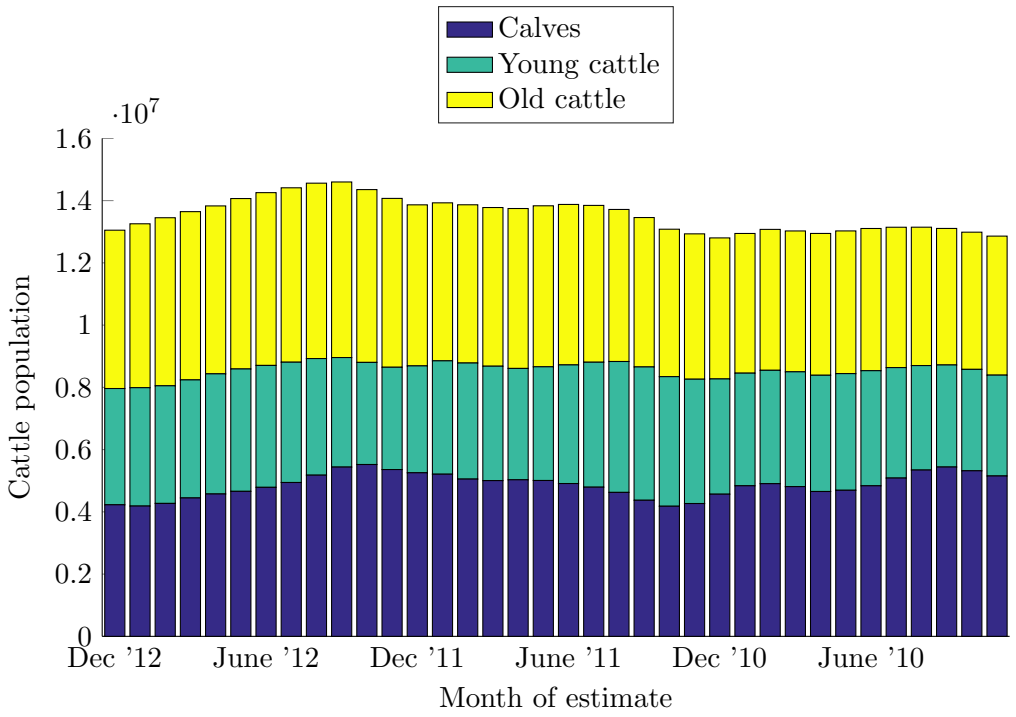

Figure 2.19: Population estimates for each from December 2012 back to January 2010. These estimates are broken down into numbers of calves (less than 12 months old), young cattle (between 12 and 24 months old), and old cattle (greater than 24 months old). The number of cattle as estimated by the Turkish Statistical Institute is also given.

estimates from the Turkish Statistical Institute (TSI)[iv], table 2.7. The TSI estimates only offer one total number for the year but the date of this estimate is not known. We can see that our estimates for August 2012 matches up with the TSI estimate of 14 million cattle. While our 2012 estimates are close to the TSI estimates we appear to over estimate the cattle population in 2011 and 2010 by approximately 2 million cattle each year.

The TSI also breaks the cattle population down into age demographics defined as calves, young cattle and old cattle, where calves are less than 12 months old, young cattle are between 12 and 24 months old, and old cattle are greater than 24 months old. We can compare the age distributions we calculated at each month to see where the disparity between our estimates and the TSI figure lies.

Our population estimate for 2012 successfully captured the number of cattle in Turkey as recorded by the TSI. Examining the demography for the year shows that we successfully predicted the number of young cattle in the country with the

[iv] `http://www.turkstat.gov.tr`

age	TSI	min	max
	2012		
calves	3,396,668	4,185,861	5,521,545
young cattle	3,623,197	3,280,538	3,932,011
old cattle	6,895,047	5,085,384	5,640,158
total	13,914,912	13,049,982	14,599,054
	2011		
calves	3,152,597	4,180,656	5,256,592
young cattle	3,177,779	3,433,763	4,282,091
old cattle	6,055,961	4,669,373	5,173,728
total	12,386,337	12,932,204	13,928,826
	2010		
calves	2,796,966	4,568,113	5,441,021
young cattle	3,028,406	3,240,515	3,748,983
old cattle	5,544,428	4,385,956	4,58,3871
total	11,369,800	12,802,074	13,146,477

Table 2.7: Comparison of the number of cattle in each age demographic from the Turkish Statistical Institute with the minimum and maximum from our estimates.

TSI estimate of 3.6 million lying with the range of 3.2 - 3.9 million predicted by our model. However discrepancies occur in the other two age demographics; our predicted range for the number of calves is 4.1-5.5 million with the TSI estimate falling well below this at 3.4 million while our predicted range for the number of old cattle is 5.1-5.6 million which falls significantly short of the TSI number of 6.9 million. These results are given in table 2.7 along with the results for 2011 and 2010.

Our estimates ultimately derive from the lifespan distribution given in figure 2.17. The peak of this distribution occurs at 19 months as we are successfully predicting the numbers of young cattle defined as being between 12 and 24 months old this would suggest that the peak of this distribution is in the correct place.

The underestimation in the number of older cattle could be due to death records being less accurate the further back in time we go, hence leading to a the number of cattle in the tails of the distribution being less than it should be. This in turn would lead to a re-normalisation of the distribution which may account for the overestimation in the number of calves.

While these approaches to choosing which cattle should still be alive are useful in providing us with estimates of the cattle population they do not provide dates on which cattle should have died. This data must be generated in order to simulate past outbreaks successfully. In chapter 5 we explicitly model outbreaks in

two Turkish farms. In that chapter we discuss methods for extrapolating dates of death based on movement data and lifespan distributions specific to these farms. For national scale modelling we would extend these techniques to each farm.

We can also attempt to remove the bias of first of the month reporting; this can be done in a number of ways. One method would be to reassign the birth days of a proportion of the cattle born on the first of the month to days in the following month. For cattle born on the first day of the year we can redistribute a proportion of their birth dates based on the average numbers of births per month. Care must be taken to ensure that cattle are not moved or die before their new birth date.

A second method would be to use the date of a cow's first move and then work backwards based on the distribution of ages at which cattle have their first move. This method would only apply to those cattle with a recorded move but it can also be used to assign a birth date to the few cattle with no birth record.

2.6 Outbreak data

As well as the cattle level data discussed above we were also provided with FMD outbreak data. This data consisted of 9,282 farm level outbreaks from January 2001 to July 2012. The locations of these outbreaks were referenced by farm name rather than a farm identity number as provided in the TurkVet movement data. These two data sets had to be reconciled by string matching in the same way as the geo-location data previously.

The majority of the outbreaks have been serotyped as type O, type A or Asia-1, though 2,847 have not been serotyped. A histogram of the number of outbreaks per month is shown in figure 2.20. From this we can see that type A and O have been the dominant strains of FMD in Turkey for the past decade. Asia-1 is not persistent and had died out in Turkey by 2002, however the strain has re-emerged in 2012.

An analysis of FMD outbreaks in Turkey from 1990 to 2002 was carried out by Gilbert et al. [2005]. The authors used the number of outbreaks in each province per year to check for spatial and temporal correlations for each of the three serotypes of FMD circulating in Turkey. The authors found that all three serotypes persisted spatially in certain provinces namely Ankara and Ezurum; as we have seen from the movement maps above Ankara and Ezurum are key areas in the cattle movement trade.

Gilbert et al. [2005] also built a meta-population model to test various possible covariates of disease status in the provinces. The authors did not have move-

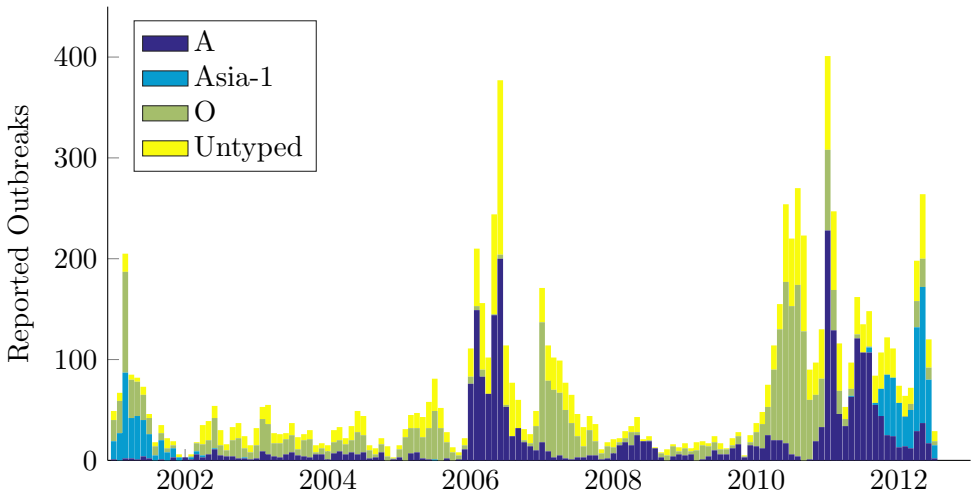

Figure 2.20: Total number of outbreaks per month from 2008 until 2011. The monthly number of outbreaks are divided into the number of each serotype.

ment data available, so instead, they used meat production-demand discrepancy as a surrogate for cattle movement. Cattle numbers, proximity to infected provinces, and meat production-demand discrepancy were found to be associated with FMD occurrence in a province. We do not repeat the analysis of Gilbert et al. [2005] here and instead focus our analysis on the movement of cattle between farms, districts and provinces.

2.7 Network data

As mentioned in the introduction livestock movement data can be used to create a network. The nodes for this network will be farms and the edges will be movements between these farms. We will create a weighted and directed network described by an adjacency matrix such that

$$A_{ij} = w_{ij} \tag{2.2}$$

where w_{ij} is the weight (which may be zero) of a link from j to i. For the calculation of some network statistics it may be necessary to force the network to be symmetric $A_{ij} = A_{ji}$.

As with the analysis of the raw numbers of movements it is important to consider different temporal resolutions for the data. For Turkey we have daily movement data for 6 years, and could in theory, create 2,192 daily networks. Analysing these daily networks would be tedious and would probably not bring much insight. On the other extreme is to aggregate all movements and create one network, this network

would be easy to analyse but would be unable to highlight seasonal trends.

We must also decide how to weight the network. There are three options; frequency weighted - each edge is weighted by the number of batches of cattle that move between two farms, cattle weighted - each edge is weighted by the number of cattle that move between two farms, and binary - each edge is one or zero determined by any movement of cattle between two farm.

From a data collection point of view it is easier to record the number of batches rather than the number of cattle moved. From a modelling point of view it is unclear which is the correct approach to take. If one believes that increasing the number of animals in a batch does not increase the potential for disease transmission much above the baseline expected by a single movement than it is legitimate to focus on the number of batches rather than the number of cattle.

2.7.1 Network statistics

We will now describe a collection of network statistics traditionally applied to the analysis of livestock networks. As we introduce each statistic we will discuss its implications on disease spread and calculate its value for the Turkish cattle network. Further discussion of all of these statistics can be found in the book by Newman [2010].

Degree distribution

The degree k_i of node i in a network is the number of edges connected to it. For a directed graph there will be two degree measures: an in-degree, k_i^{in} and an out-degree, k_i^{out}. The degree of each node can easily be calculated by summing rows or columns from the adjacency matrix

$$k_i^{\text{in}} = \sum_{j=1}^{N} A_{ij}, \quad k_j^{\text{out}} = \sum_{i}^{N} A_{ij}. \tag{2.3}$$

If the network is symmetric then $k_i^{\text{in}} = k_i^{\text{out}} = k_i$. The total number of edges in a directed network is

$$m = \sum_{i=1}^{N} k_i^{\text{in}} = \sum_{j}^{N} k_j^{\text{out}} = \sum_{ij} A_{ij}. \tag{2.4}$$

The degree distribution of a network is of particular interest in the study of networks. The degree distribution is used to categorise types of networks. One of the most common degree distributions is a power law degree distribution which can describe a scale-free network. Scale free networks arise when there is a very

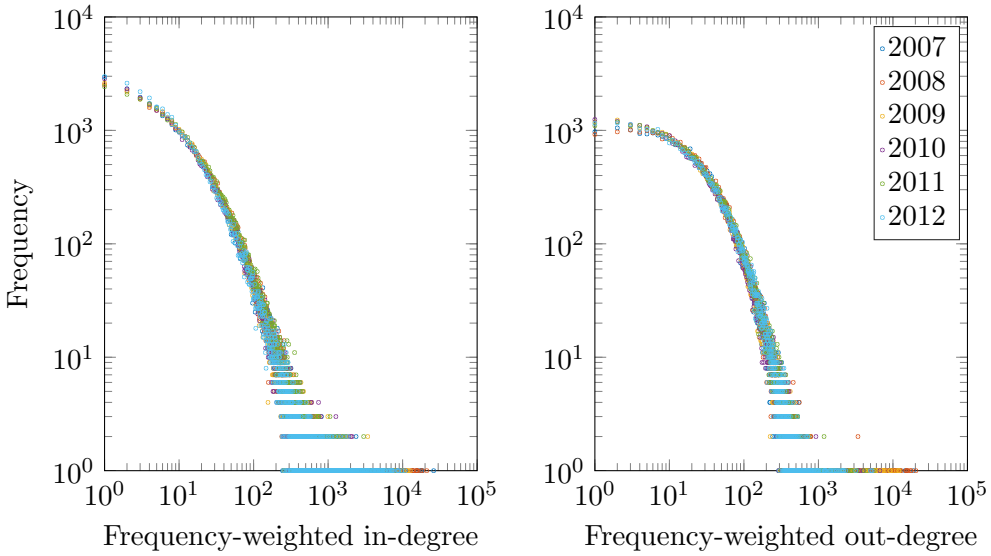

Figure 2.21: Frequency weighted in and out degree distribution from 2007 to 2012.

heterogeneous contact structure; most nodes have a low degree with a few nodes having a very high degree. Examples of such networks are discussed in Albert and Barabási [2002] and include: academic citation networks, the world wide web, food webs and sexual contact networks.

In a scale free network the fraction of nodes, $P(k)$ with degree k will be described by

$$P(k) \sim k^{-\gamma}, \qquad 2 < \gamma < 3. \tag{2.5}$$

The most famous method for reproducing networks with a power law distribution is the preferential attachment model of Barabási and Albert [1999]. Many networks are said to fit this power law distribution, most famously the world wide web. Difficulties can arise in fitting a power law to a distribution that appears scale free due to large fluctuations in the tail of the power law [Clauset et al., 2009].

The frequency weighted in and out degree distributions for the annual Turkish networks are shown in figure 2.21. These distributions demonstrate fat tails which are indicative of scale free networks. Annually there is little variation in the shape of the in-degree and out-degree distributions. There is however a consistent variation between the shape of the in-degree and out-degree distributions for a given year. The in-degree distributions have more bulk for low-degree nodes with 40% of farms having 10 or fewer in-movements while only 25% of farms have 10 or fewer out-movements.

Degree correlations and assortativity

Degree correlations are important for the spread of disease [Danon et al., 2011]. One of the most basic measures of degree correlation applied to a directed network is that of the correlation between in-degree and out-degree. We use the Pearson correlation coefficient $\rho(\alpha, \beta)$ where $\alpha, \beta \in \{\text{in}, \text{out}\}$ to quantify this correlation,

$$\rho(\alpha, \beta) = \frac{\text{covariance}(\alpha, \beta)}{\text{standard deviation}(\alpha) \times \text{standard deviation}(\beta)}. \tag{2.6}$$

High positive values of $\rho(\text{in}, \text{out})$ indicate that the in-degree and out-degree are correlated. Thus, there will be farms that are both heavily involved in the buying and selling of cattle. These farms are both likely to become infected and to spread the infection to other farms Kiss et al. [2006].

We can also examine the correlation between a farm's degree at one point in time with other time points. This will indicate if there is a shift in the temporal trading pattern. We investigate the correlations between in-degree and out-degree from 2007 to 2012. These results will populate four matrices $\mathbf{R}(\alpha, \beta)$ with entries

$$R(\alpha, \beta)_{ij} = \rho(\alpha(i), \beta(j)) \tag{2.7}$$

where $i, j \in 2007 : 2012$. We note that $\mathbf{R}(\alpha, \beta) = \mathbf{R}(\beta, \alpha)^{\mathbf{t}}$, and $\rho(\alpha(i), \alpha(i)) = 1$.

Figure 2.22 shows all four of these matrices. What is most notable is that two blocks seem to form within each matrix; $\mathbf{R}(\text{in}, \text{in})$ and $\mathbf{R}(\text{out}, \text{out})$ for the years 2007-09 are highly correlated with each other but not with the following three years. The first three years have an average $\rho(\text{in}, \text{in})$=0.98 while their correlation with the following 3 years drops to 0.57 on average.

Within the same year the average correlation between in-degree and out-degree is $R(\text{in}), \text{out})_{ii} = 0.76$. This lies within the range of values that [Kiss et al., 2006] reported for the sheep movement network on Great Britain.

The correlations calculated above describe individual nodes. We will now turn our attention to higher level correlations in the interaction of nodes with their immediate neighbours.

If similar nodes in a network connect to each other more frequently than they connect to dissimilar nodes we describe that network as being assortative; if the converse is true and similar nodes were connected to dissimilar nodes more often than to similar then the network is disassortative.

In social networks we may describe people as similar categorically by race or job title or we can describe their similarity numerically by age or income. These

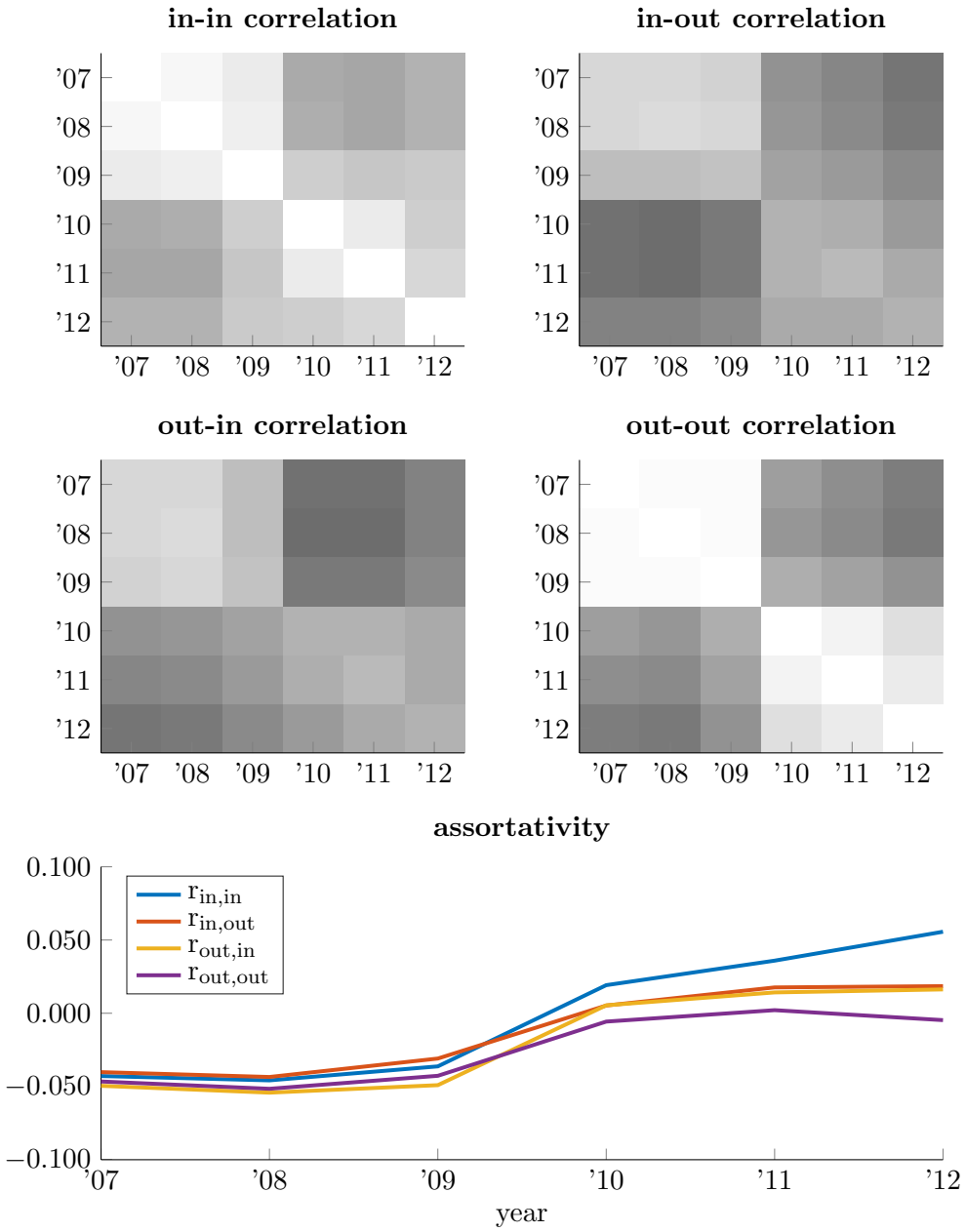

Figure 2.22: The four color maps indicated the correlations between in-degree and out-degree correlations across the years 2007 to 2012 using annual frequency weighted networks. The scale ranges from 0 to 1 with white indicating a 1. The lower plot shows the assortativity coefficients as defined by equation 2.8 for each year.

measures of similarity require an additional layer of information separate to the network. One way we can classify nodes as similar that is entirely network dependent is by using the node's degree. Thus we can call a network assortative if nodes with high degree connect to other high degree nodes and nodes with low degree connect to other low degree nodes.

This notion of degree assortativity is particularly relevant to the study of networks in an epidemiological context as it has been shown that epidemics can spread quicker on assortatively mixed networks while control is easier to implement on a dissasortatively mixed network [Kiss et al., 2006].

For an undirected network degree assortativity is measured using the Pearson correlation function. The assortativity value ranges from -1 for a totally dissortative network to +1 for a totally assortative network. For directed networks this concept was extended by Foster et al. [2010] who introduced four measure of degree assortativity:

$r_{\textbf{in-in}}$ the assortativity between the in-degree of the incident node and the in-degree of the target node

$r_{\textbf{in-out}}$ the assortativity between the in-degree of the incident node and the out-degree of the target node

$r_{\textbf{out-out}}$ the assortativity between the out-degree of the incident node and the out-degree of the target node

$r_{\textbf{out-in}}$ the assortativity between the out-degree of the incident node and the in-degree of the target node.

By letting $\alpha, \beta \in \{\text{in}, \text{out}\}$ where k_i^α and l_i^β are the α-degree and the β-degree of the source and target nodes for edge i respectively, we can define assortativity measures using Pearson's correlation coefficient

$$r(\alpha, \beta) = \frac{M^{-1} \sum_i (k_i^\alpha - \bar{k}^\alpha)(l_i^\beta - \bar{l}^\beta)}{\sqrt{M^{-1} \sum_i (k_i^\alpha - \bar{k}^\alpha)^2} \sqrt{M^{-1} \sum_i (k_i^\beta - \bar{k}^\beta)^2}} \tag{2.8}$$

where M is the number of edges in the network and $\bar{k}^\alpha = M^{-1} \sum_i l_i^\alpha$.

The assortativity measures are plotted in the lower half of figure 2.22. As with the values of ρ the values for the assortativity measures undergo change with time from the first three years considered to the second three. Initially each of the four measures begins slightly negative and within a tight range. From 2010 the

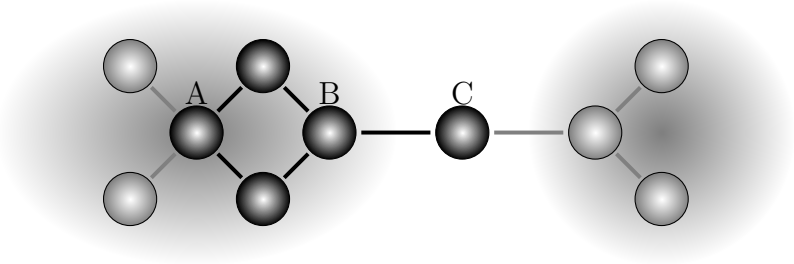

Figure 2.23: There are two shortest paths from A to C both passing through B along the black edges and nodes this will increase B's betweenness centrality (as described by equation 2.11) by $\frac{1}{2} + \frac{1}{2} = 1$. Node C acts as a bridging node between the two groups of nodes that would otherwise be disconnected. Many of the shortest paths in the network will pass through C which will give it a high betweenness centrality.

measures become positive and the range between them widens, particularly between $r(\text{in}, \text{in})$ and $r(\text{out}, \text{out})$.

Despite their temporal variation the values of these measures still lie in a reasonably narrow range $-0.06 < r(\alpha, \beta) < 0.06$ indicating the network is neither strongly assortative or dissasortative. However, the switch from negative to positive values could be indicative of a structural change in the network that may be result in disease outbreaks being more difficult to control.

Paths in the network

A path in the network is any ordered set of nodes such that every consecutive node in the set is connected by an edge. Edge weight is not taken into account but the direction can be. The average path length can be calculated as can the distribution of path lengths. There will not necessarily be a path between every pair of nodes in the network but for those that have a path the shortest path between the pair is also of interest. The shortest path is not necessarily unique as illustrated in figure 2.23.

Connected components

If a path exists for every pair of nodes in the network the network is said to be connected. When this is not the case the network is disconnected, but groups of nodes that form connected sub-networks are called connected components. For directed graphs there will be two types of connected components; strongly connected components (SCCs) and weakly connected components (WCCs). A group of nodes form a strongly connected component if paths that obey edge direction exists between

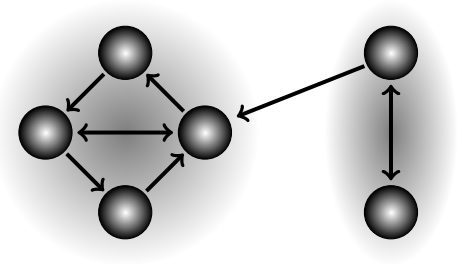

Figure 2.24: This network has two strongly connected components. If edge direction was ignored the entire network would form a connected component; in the directed case it forms two weakly connected components.

every node in that group. A group of nodes form a weakly connected component if paths exist between every node in that group regardless of edge direction. An example of connected components for a directed network is shown in figure 2.24.

If the largest strongly connected component contains a sizeable fraction of the nodes of the networks it is known as the giant strongly connected component (GSCC). For the Turkish network a GSCC component exists. We will compute all of the network measures for nodes contained within the GSCC. This increases computational speed and removes nodes with low or sometimes zero degree which are poorly connected and would have negligible influence on disease dynamics.

We computed the size of the GSCC for each month from January 2007 to December 2012. As seen in figure 2.25 there is a clear biannual trend. The first peak of each year occurs in spring when new young cattle are born and moved to pasture. The second peak corresponds to the Kurban festival.

Similar seasonal trends appear in the French cattle network [Dutta et al., 2014] and the British cattle network Vernon [2011], both of which have a peak in the Spring and a second peak near the end of the year.

The number of movements each month mirrors the size of the GSCC and is also shown in figure 2.25. Increased movement during these biannual peaks leads to greater opportunity for disease spread. The size of the GSCC acts as a lower bound on the maximum outbreak size.

Each month many farms have no movements at all. The number of inactive farms is inversely correlated to the size of GSCC and can be quite large, figure 2.25. Unless otherwise stated we will only consider statistics relating to the farms within the GSCC so that farms that are inactive do not affect average statistics.

The average binary degree within the GSCC is plotted in figure 2.26. The biannual trend exhibited by the GSCC size is again apparent in the average degree. Following the analysis of Kiss et al. [2006] on the UK sheep network we

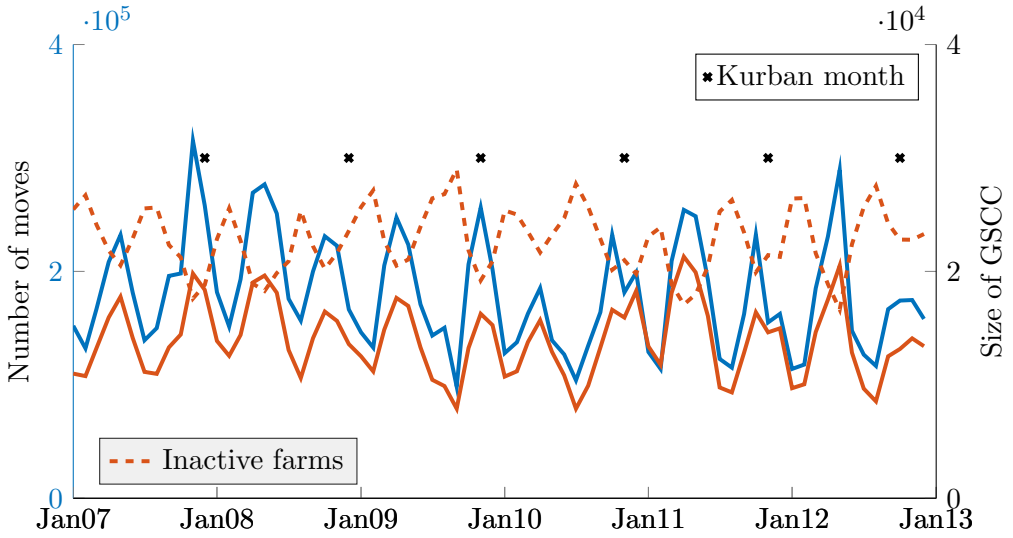

Figure 2.25: The number of moves in the full network and the size of the giant strongly connected component (GSCC) are plotted for each of the frequency weighted monthly networks from January 2007 to December 2012. The number of inactive farm is also shown. The months at which the Kurban festival occurred in each year are shown for reference.

also present the proportion of links that are bi-directional (that is are reciprocated between farms). Disease spread is hindered by the presence of a large proportion of bi-directional links.

The range of the proportion of bi-directional links is rather narrow from 0.06 to 0.16 when compared to the UK sheep network which ranged from 0.1 to 0.6 but is close to that observed for the French cattle network which varied from 0.13 to 0.16. The inverse correlation between this proportion and the average degree within the GSCC observed by Kiss et al. [2006] is also observed here.

Notable too is the reduction in the proportion of bi-directional links between the first three years and the second three; another sign of a structural change in the network. One would expect that a decreased proportion of bi-directional links could lead to an increase in clustering in the network.

Clustering

If a group of nodes in a network are highly connected to each other we speak of them as forming a cluster. The quantification of clustering comes from counting the number of triangles in a network. Triangles are important as they indicate transitivity between nodes. There are two types of clustering; global clustering and local clustering.

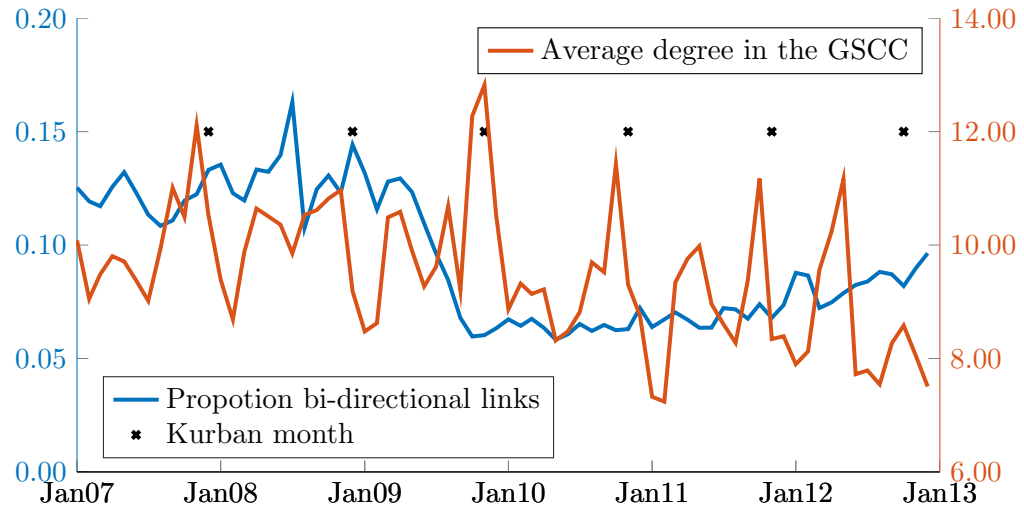

Figure 2.26: The proportion of links within the GSCC that are bi-directional are shown for the binary monthly networks from January 2007 to December 2012. The months at which the Kurban festival occurred in each year are shown for reference.

Global clustering ϕ is a straightforward count of the triangles in the network divided by the number of connected triplets in the network. The adjacency matrix for a symmetric binary network when raised to a power l conveniently counts paths of length l, for example $A^3_{ij} = 4$ means there are 4 paths of length 3 between nodes i and j. The diagonal elements of A^l_{ii} correspond to the number of paths of length l that start and finish at i. Hence,

$$\phi = \frac{\text{trace}(\mathbf{A}^3)}{||\mathbf{A^2}|| - \text{trace}(\mathbf{A^2})} \tag{2.9}$$

where the trace($\mathbf{A}$) is the sum of the diagonal elements of $\mathbf{A}$ and $||\mathbf{A}||$ is twice the total number of edges in $\mathbf{A}$. This formula is true for undirected graphs. For directed graphs it will only count cycles as triangles.

As its name suggests the global clustering defines how clustered the entire network is. At a local level we can also define a clustering coefficient that counts how often the neighbours of a node are also neighbours themselves. This measure was introduced by Watts and Strogatz [1998] to describe small-world networks. The local clustering coefficient for node i is

$$C_i = \frac{\text{all triangles formed by } i}{\text{all possible triangles } i \text{ could form}}. \tag{2.10}$$

an average can be taken over all nodes to give an average local clustering coefficient,

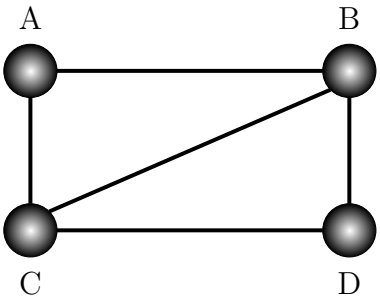

Figure 2.27: Node C has three neighbours A, B and D; C is connected to B and B is connected to D while A and D are not connected. This gives C a local clustering coefficient of $C_C = \frac{2}{3}$. The C_i's for the other nodes are $C_A = C_D = 1$,$C_B = C_C = \frac{2}{3}$. The average local clustering coefficient $C = \frac{5}{6}$ which is not equal to the global clustering $\phi = \frac{6}{10} = \frac{2}{3}$. If we were to reweight every C_i by $d_i(d_i - 1)$ and then average then $C = \phi$.

C. This formula is easily extended to directed networks by considering directed triangles, 8 of these are possible for a triplet of nodes.

The average local clustering coefficient, C is in general not equal to ϕ. For some small networks like the one shown in figure 2.27 they can be the same but in general they are not. C tends to be dominated by low degree nodes as they have small denominators and thus may give a poor representation of the clustering of the network [Newman, 2010].

The effect of clustering on epidemics has been discussed in [Keeling, 2005b]. Here the author showed that the structure of random networks with the same average degree has a bearing on epidemic dynamics. Initially higher clustering can lead to increases in the calculated R_0 however as the pool of susceptible nodes is depleted through further generations networks with higher ϕ show a lower asymptotic value for R_0.

Box plots of the local clustering coefficients for annual networks are plotted in figure 2.28. The range is quite large with many farms having a C_i of 0 or 1. Farms at these extremes will have low degree. The median value and the interquartile range for C_i for the first three years are higher than for the second three years.

At a monthly level the transition in the C_i values is again observed, figure 2.29. As mentioned above the mean value of C_i has been proposed as a measure of global clustering [Watts and Strogatz, 1998], and is plotted on top of the box plots for C_i. However, as discussed by Newman [2010], this average can be dominated by low degree nodes and may not reflect the macroscopic behaviour of the network.

To gain a better understanding of the global behaviour of the network we

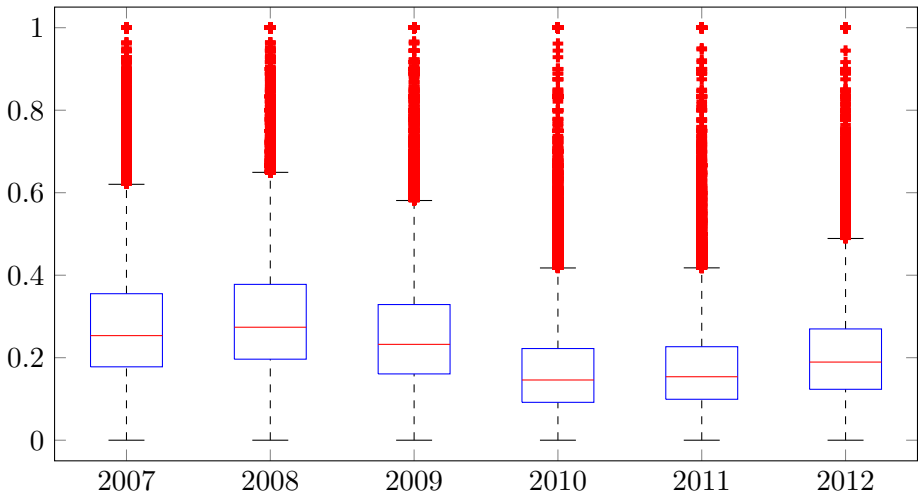

Figure 2.28: Box plots showing the local clustering coefficients for farms in the GSCC for each annually constructed network from 2007 to 2012.

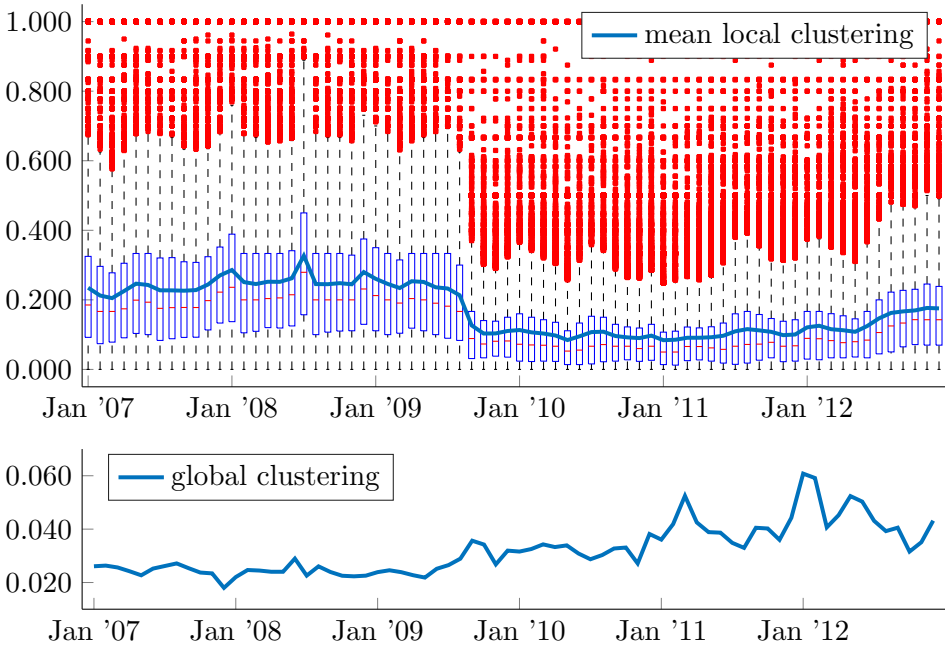

Figure 2.29: Box plots showing the local clustering coefficients for farms in the GSCC for each monthly constructed network from 2007 to 2012. The mean local clustering coefficient is plotted for each GSCC as is the global clustering coefficient.

also calculate ϕ for each of the monthly networks, figure 2.29. The range of ϕ is small (between 0.02 and 0.06) but again a change is apparent in the latter three years examined. It is interesting to note that if the network were undirected then weighting each C_i by $d_i(d_i - 1)$ before averaging we can recover ϕ.

Despite the average of the local clustering coefficient not necessarily being suitable for a global analysis of the network; the distribution of individual clustering coefficients still reveals important local information. Nodes with with low C_i have neighbours who do not share many edges in common, node i could therefore control the flow of information to its neighbours. Thus, C_i can act as a measure of node centrality.

Betweenness centrality

A large portion of research into networks is dedicated to finding the most important nodes in a network [Newman, 2010]. Importance is obviously a subjective idea and as such different centrality metrics have been proposed according to how importance has been defined. We have already come across node degree which, if importance was defined in terms of the number of connections a node has, would act as the degree centrality of the network.

A more refined measure of importance might be to consider not the number of connections a node has but how well the node facilitates connections between groups of other nodes. In a livestock disease context the important nodes may be the ones that make relatively few trades but whose trades are between two groups of farms that otherwise might be disconnected. Betweenness centrality $b(i)$ attempts to measure this type of importance by counting how often a node lies on the shortest path between every other pair of nodes,

$$b(i) = \sum_{i \neq j \neq k} \frac{\sigma(i)_{jk}}{\sigma_{jk}} \tag{2.11}$$

where $\sigma(i)_{jk}$ is the number of shortest paths from node j to node k passing through node i and σ_{jk} is the total number of shortest paths from j to k.

We note that equation 2.11 is not just a sum over the number of shortest paths from j to k that pass through i but that it is normalised by the total number of paths from j to k. In this way if there are two shortest paths between a pair of nodes then each gets a weight of $\frac{1}{2}$. Furthermore, if there are three shortest paths between j and k and two of them pass through i then this will contribute $\frac{2}{3}$ to $b(i)$. An example of betweenness centrality for a network without edge direction is given

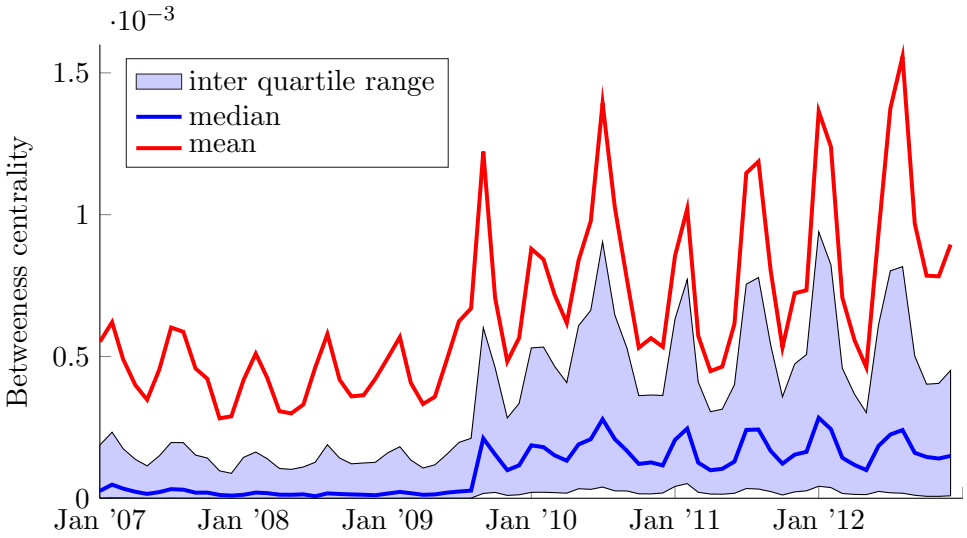

Figure 2.30: The normalised betweenness centrality for farms in the GSCC for the cattle networks constructed monthly. The median and mean values are plotted as is the interquartile range. For clarity values outside this range have not been plotted.

in figure 2.23.

We can also normalise $b(i)$ for comparison with other networks. To do this we divide by $(N-1) \times (N-2)$ which is the highest number of shortest paths that could pass through one node in a directed network.

The monthly betweenness centrality for Turkey is plotted in figure 2.30. There is once again a change in the values between the first and second set of years, with the betweenness centrality increasing. The mean value is outside the interquartile range indicating that a few nodes with high centrality dominate the average value.

The betweenness centrality offers a more global perspective of a nodes centrality than local clustering though a slight inverse correlation between the two is observed. There is a seasonal trend in $b(i)$ that was not evident in C_i. Many nodes having high betweenness centrality can facilitate fast spread of infection through the network but may make control efforts easier could be targeted to these nodes.

2.7.2 Discussion of network statistics

A definite change in some of the calculated network properties is evident between the years 2006-09 and 2010-12; while the component size and number of edges have not changed, reciprocity has decreased meaning edges are less likely to be shared. We see that the assortativity has increased implying that previously edges were

shared between small and large farms but now farms are tending to trade slightly more with farms of their own size. How these formerly reciprocal edges are now being placed will determine the network statistics.

The edges that were previously being reciprocated between the large and small farms are now being used to connect large farms to other large farms and small farms to other small farms. This could be the cause of the reduction in clustering coefficients for small farms as collectively small farms share few edges between them and are less likely to form triangles. Conversely, large farms are now being connected more often possibly completely cycles and causing an increase in global clustering. Large farms connecting more regularly to each other will place more of these farms on the shortest paths across the network and will increase the betweenness centrality.

The statistics calculated above are at a node level and effectively attempt to establish the importance of nodes via some centrality measure. In chapter 4 we will introduce the concept of community structure which looks beyond these measures and attempts to group nodes which share edges together to give a higher order structure. We will implement this analysis for the cattle networks of both the UK and Turkey.

Chapter 3

Modelling livestock disease with partial data

3.1 Introduction

In the UK, an annual livestock census records the location and species composition of all livestock farms. Births, deaths and movements of animals are recorded for individual cattle via the Cattle Tracing Scheme (CTS) and for batches of other livestock via the Animal Movements Licence Scheme (AMLS). Such data have driven the development of sophisticated models to capture and predict the spread of livestock diseases such as FMD [Ferguson et al., 2001a; Green et al., 2006; Ortiz-Pelaez et al., 2006; Tildesley et al., 2008], bovine tuberculosis [Brooks-Pollock and Keeling, 2009; Green et al., 2008] and E-coli [Zhang et al., 2010].

However, many countries around the world do not routinely collect farm-level data, or they are not readily available for research owing to issues regarding privacy. For example, in the USA, the National Agricultural Statistics Service (NASS) carries out an agricultural census every 5 years. In order to preserve anonymity for farmers, all data are aggregated at the county level and therefore precise locations of livestock farms are unknown. Furthermore, movement data are held at the individual state level and there is no requirement for livestock movements to be recorded unless movements are out of state [Buhnerkempe et al., 2013]. In the UK, the poultry industry infers movements between holdings using targeted sampling of premises based upon their function and size. This method predominantly targets large farms and therefore does not accurately capture the demographic characteristics of the underlying farm population [Nickbakhsh et al., 2011]. Therefore, it is important to understand the ability of models to predict the potential for disease spread through

livestock movements when only a partial sample of the network is available.

Partial network data is a well-known problem and has been studied extensively in social sciences and other fields such as epidemiology in human and livestock diseases. While the amount of data available to modellers is increasing, so too are privacy concerns. In order to predict the risk of disease spread in humans across large spatial scales, detailed movement networks must be established. These networks can be informed using commuting and migration data available from population censuses [Danon et al., 2009; Pindolia et al., 2013; Wesolowski et al., 2013]. These data capture long term trends but may not be appropriate at predicting movements over a shorter time scale and therefore can be complemented by the inclusion of other information such as mobile phone data. Mobile phone records track locations and times that individuals make and receive calls and therefore can act as a proxy for shorter scale movement patterns [González et al., 2008; Wesolowski et al., 2013]. Whilst full access to these data sets is not readily available, previous work indicates that partial samples may be sufficient to accurately predict the risk associated with disease spread across these networks [Pindolia et al., 2013; Tizzoni et al., 2014].

In situations where only partial network information are available, it may be necessary to reconstruct the network. Different approaches can be applied to construct contact networks. The most basic methods involve random sampling of nodes, i.e. individuals or farms [Stumpf et al., 2005] or random sampling of edges (i.e. links between nodes). However, it may be possible to capture the key properties of a network more efficiently using an approach such as snowball sampling. Snowball sampling is typically used in situations where the target population is small and hard to find. A number of sampled individuals from the target population are asked to nominate a set number of other people from the target population [Goodman, 1961]. This method has been used previously to identify networks of sexual contacts for HIV positive individuals [Kendall et al., 2008]. In the case of livestock, when there may be knowledge regarding the size of farms or number of aggregated movements from a farm (per year), a targeted sampling approach could be used where larger farms or farms with the highest number of movements are sampled. When specific network characteristics, such as age-structures, are included in the construction of networks this improves the quality of the constructed network [Fumanelli et al., 2012]. Previous studies have shown that subsets of networks are not always representative of the whole network in some cases (e.g. scale-free networks [Stumpf et al., 2005]). Therefore, caution is needed when networks are constructed with partial data.

In previous work Tildesley et al. [2010] demonstrated that even in the absence

of precise locations of farms, accurate predictions of the impact of interventions is possible. We aim to develop an understanding of the predictive power of mathematical models when only a subset of the network information is available. We develop a model to simulate the spread of a rapidly spreading disease such as FMD through the UK cattle movement network. Mathematical models have previously played a key role in determining the risk of disease spread through networks of livestock movements for diseases such as FMD [Buhnerkempe et al., 2013; Green et al., 2006; Kiss et al., 2006; Ortiz-Pelaez et al., 2006], BTB [Brooks-Pollock and Keeling, 2009; Brooks-Pollock et al., 2014; Green et al., 2008] and Bluetongue Virus [Ensoy et al., 2013; Szmaragd et al., 2009]. The aim of this chapter is to investigate the ability of such models to provide policy advice in countries where only partial information regarding livestock movements is available.

We compare four imperfect data types: random sampling of movements (weighted edges), random sampling of farms (nodes), snowball sampling [Kendall et al., 2008; Kolaczyk, 2009] of farms and targeted sampling of farms. If appropriate we then scale the sampled networks up so that the original number of movements are used for the epidemic simulations. In the UK, selling and buying of livestock often takes place through livestock markets. Previous work suggests that these markets played a substantial role during the 2001 FMD outbreak [Gibbens et al., 2001]. As animals from different farms are kept in close proximity there is a risk of disease transmission between batches of animals resulting in spread of infection to multiple farms. Moreover, it is known that movements from markets cover a large geographical area [Robinson and Christley, 2007]. Therefore, we investigate the potential role of markets in disease transmission between farms.

This study will be highly informative for countries where livestock movement data are not routinely available. The outputs of this work will provide guidance to livestock industries around the world regarding both how much data are required to predict spread of disease and how to target data collection should it not be possible to record all livestock movements.

3.2 Materials and methods

In this chapter we utilize data from the 2010 CTS database (provided by the Department for the Environment, Food and Rural Affairs, Defra, via the Animal and Plant Health Agency, APHA) for Great Britain. If multiple animals were moved on the same day from one farm to another this was treated as one movement; markets were initially not explicitly included. Slaughterhouses were considered as sinks and

therefore movements to slaughterhouses were ignored even when going through a livestock market. In total there were 70,243 farms and 327 markets in our dataset with 856,454 movements in total. A total of 635,016 movements passed through markets, with 47,692 farms using cattle markets at least once during 2010.

Four methods of sampling from this database are implemented and compared here. A directed weighted-static adjacency matrix A [Vernon and Keeling, 2009] was constructed for each set of sampled data, in which nodes represent farms and edges represent (directed) cattle movements. An edge a_{ij} is defined to be non-zero if cattle are moved from farm j to farm i during the year. The weight of the edge represents the frequency of movements from farm j to farm i in 2010, i.e. the total number of days on which movements occurred divided by 365.

3.2.1 Movement sampling

For random movement sampling RMS, we list the recorded movements and randomly sample from this list. The depleted network is then built from the remaining movements and the resultant network is rescaled such that the total weight of the rebuilt network is equal to that of the original network:

$$A^{\text{rebuilt}} = \frac{W^{\text{original}}}{W^{\text{depleted}}} A^{\text{depleted}} \tag{3.1}$$

where $W^{\text{original}} = \sum_{ij} a_{ij}^{\text{original}}$ and $W^{\text{depleted}} = \sum_{ij} a_{ij}^{\text{RMS}}$. This method explicitly depends on knowledge of the total weight of the network.

3.2.2 Node sampling

In the node sampling schemes a "sampled" node has all its edges sampled. A node is said to be "captured" if it is connected to a sampled node but it has not been sampled itself. For all of these schemes we assume that the total number of nodes N, in the network is known. We sample a set S of N_S nodes and capture N_C nodes, the N_S sampled nodes plus their connected non-sampled neighbours. This method will therefore preserve the degree of the initially sampled nodes N_S but for the remaining nodes that are captured, only the edges that link them to the N_S nodes will be recorded.

We consider three methods of node sampling in this chapter. In the first method, we use random node sampling (RNS) such that all movements from a certain percentage of nodes are selected. A more advanced form of node sampling is snowball sampling (SBS) [Kolaczyk, 2009]. In this method, an initial set of nodes are

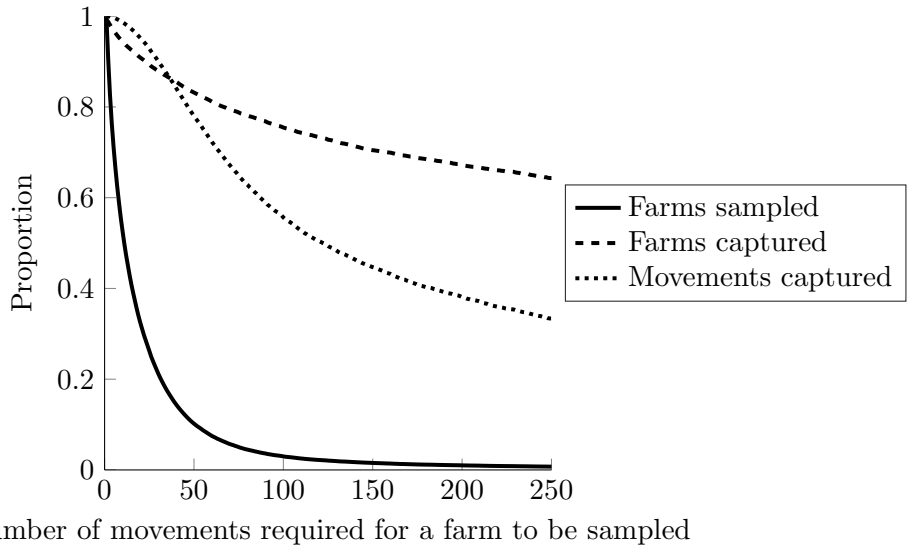

Figure 3.1: Plot showing the proportion of farms with at least x number of movements. The proportion of farms and movements captured when the targeted node sampling scheme is used to sample the farms with at least x number of movements is also shown.

sampled at random. At the next stage the nodes captured by the initial sampled nodes are in turn sampled. This process can continue until all nodes have been sampled. Here, we consider second order snowball sampling, such that an initial set of nodes are sampled and the nodes that this set of nodes are connected to are also sampled.

The final node sampling scheme considered is targeted node sampling (TNS). For TNS we sample specific nodes based on certain criteria. In this case we chose the weighted-degree of the node. All nodes having a weighted-degree of at least x are sampled and their neighbours are captured. Figure 3.1 shows the proportion of farms sampled for a specific x, the proportion of farms and movements captured at that x is also shown. A graphical depiction of the three node sampling network schemes is shown in figure 3.2.

We firstly considered rescaling the network formed by the node sampling methods in a similar way to that for the RMS method. We used the average weighted-degree of the sampled nodes

$$\langle w^{\text{sampled}} \rangle = \frac{1}{N_S} \sum_{i \in S} \sum_{j} (a_{ij} + a_{ji}) \tag{3.2}$$

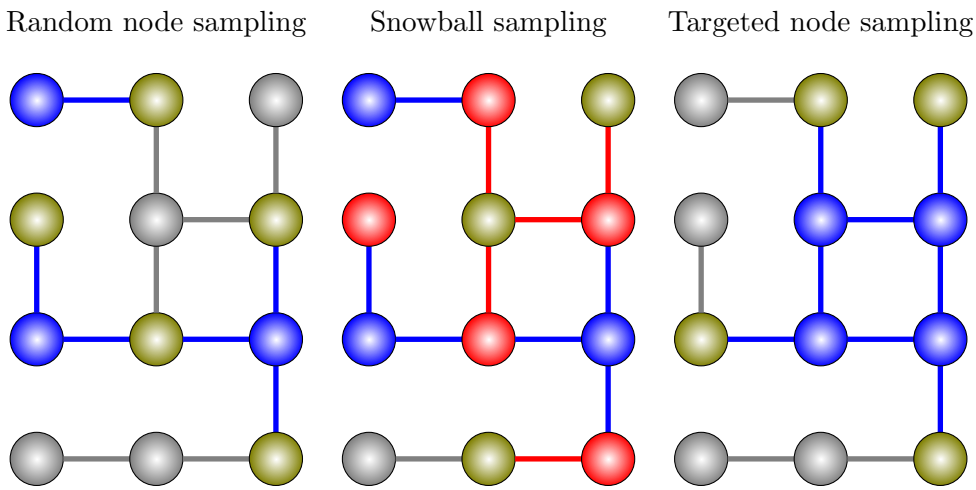

Figure 3.2: Schematic illustration showing the three sampling schemes employed. Sampled nodes and edges are in blue for initially sampled nodes and red for the case of second stage snowball sampling (centre). Nodes discovered by the sampled nodes (captured-nodes) are coloured green with their non-sampled edges in grey. Nodes which have not been sampled or discovered are coloured grey. Here second stage snowball sampling is depicted, at the third stage all green nodes would turn red, their edges would turn red and gray nodes connected to them would turn green. For targeted node sampling (right) node with a degree ≥ 3 have been sampled.

to estimate the total weight of the original network

$$\tilde{W}^{\text{original}} = \langle w^{\text{sampled}} \rangle \frac{N}{2}, \tag{3.3}$$

and rescale the network as in equation (3.1) but using the estimate for the total weight of the original network. However, as shown in supplementary figure B.2, these scaled networks result in significant overpredictions of epidemic size, particularly when small percentages of the nodes are sampled. For the remainder of this chapter, we therefore use the unscaled versions of the node sampling methods.

3.2.3 Network statistics

The properties of the underlying network may have a significant effect on epidemic dynamics [Keeling, 2005b]. We therefore consider how network properties change as less data are utilised by the various sampling schemes.

We first consider the number of strongly connected components of the network. A subset of nodes forms a strongly connected component if each of the nodes can connect to each other node by following a path which preserves edge direction. If the largest of these components is of the same order as the complete network

it is known as the giant strongly connected component (GSCC) and gives a lower bound to the maximum size of an outbreak on the network if the disease is perfectly transmissible [Kiss et al., 2006]. In addition, we explore the impact of the sampling schemes upon the mean and standard deviation of the weighted w and unweighted k degree of nodes in the GSCC.

Assortativity [Foster et al., 2010; Newman, 2010] is the tendency for similar nodes to connect to each other in a network, and is measured using the Pearson correlation function for the node attribute under inspection, generally the node degree. The assortativity value ranges from -1 for a totally dissortative network to +1 for a totally assortative network. If the correlation is zero, there is no tendency for nodes with a similar degree to connect or actively avoid each other. In the case of a directed network there are four types of degree correlation to consider:

$r_{\textbf{in-in}}$ the correlation between the in-degree of the incident node and the in-degree of the target node

$r_{\textbf{in-out}}$ the correlation between the in-degree of the incident node and the out-degree of the target node

$r_{\textbf{out-out}}$ the correlation between the out-degree of the incident node and the out-degree of the target node

$r_{\textbf{out-in}}$ the correlation between the out-degree of the incident node and the in-degree of the target node.

For a full and detailed explanation of these measures and a discussion on their implications see Foster et al. [2010].

These statistics are averaged over 1000 realisations of the network for each sampling method. The diameter of the GSCC was also measured but, owing to extensive computational time, this was only calculated for a single realisation of the network. The diameter of a network is the length of the longest shortest path across the network [Newman, 2010]. As well as the network statistics mentioned previously above we also explore the number of nodes and edges captured by the various sampling schemes as the percentage of sampled data varies.

3.2.4 Comparison of epidemic predictions

A stochastic SIR model (susceptible - infectious - recovered) was used to investigate epidemic behaviour on the livestock network. The probability of farm i becoming

infected is defined as

$$\lambda_i = 1 - \exp\left(-\beta \sum_i a_{ij} I_j\right) \tag{3.4}$$

where $I_i = 1$ if farm i is infected and zero otherwise, β is the transmission rate. Infected farms recover after a period T and cannot be reinfected.

We aim to investigate spread of relatively 'fast moving' diseases in the absence of movement restrictions such as foot-and-mouth disease. We make the assumption that transmission of infection to a farm results in all animals on that farm moving into the infectious class. Given this assumption, the risk of infection between any pair of nodes in the movement network is based upon the number of movements between them rather than the number of animals moved. In order to investigate the impact of epidemiological parameters upon model predictions, we explore a range of values for the transmission parameter and the infectious period, such that $\beta = 1, 2, 5, 10$ and $T = 7, 14, 21, 28$ days.

After reconstructing the movement network, epidemics were seeded randomly in either Cumbria, Aberdeenshire or Devon. These three counties have a high number of cattle farms and livestock movements and therefore epidemics starting in these counties are more likely to produce a high number of cases than in other parts of the UK. Cumbria and Devon were also two major hotspots of infection during the UK FMD outbreak in 2001[Keeling et al., 2001]. A random source farm in each county was infected initially for each simulation and we investigated the predicted final epidemic size, duration, peak size and the model prediction of the geographic spread of disease.

A thousand networks were created for each sampling scheme and of these, one hundred were randomly selected for simulations. Statistics are averaged over 1000 simulations that had a final epidemic size of at least ten farms. Pseudo-code for the SIR process is given in algorithm 1.

3.2.5 Livestock Markets

Markets may play a key role in amplification of disease transmission [Robinson et al., 2007]. The CTS explicitly states whether a movement went through a market and if so which market was used. This allows us to construct networks that include markets as nodes. The sampling schemes listed above can all be applied to this situation.

There is significant uncertainty regarding the level of contact of animals from different batches (farms) on a market and therefore the risk of transmission between animals during their stay on the market. For this reason, we investigate the

Algorithm 1 Standard SIR

```
1: t = 1;
2: I(k) = 1;                                  ▷ Seed farm k to initially infected
3: I_time(1) = 1                              ▷ Record time of infection
4: while ∑(I) > 0 do
5:     for i = 1 : N do                       ▷ Infection loop
6:         Calculate λ_i from equation 3.4
7:         p = RAND
8:         if λ_i < p then
9:             I(i) = 1
10:            I_time(i) = t
11:        end if
12:    end for
13:    for i = 1 : N do                       ▷ Recovery loop
14:        if I_time(i) − T = 0 then
15:            R(i) = 0
16:        end if
17:    end for
18:    t++
19: end while
```

effect of two extreme assumptions of transmission within a market (figure 3.3). In the first scenario, we assume no within-market transmission, such that infection is only transmitted between the source and the destination farm. We assume complete segregation between herds being strictly enforced (this would be equivalent to having no markets in the network).

In the second scenario, we assume no segregation and no biosecurity at a market, such that all batches that move through a market mix with one another homogeneously. In this case, we use the CTS data to determine which batches of cattle move through a market. When an infected batch moves to a market, that market becomes infected and we then assume that infection can be transmitted to all possible destination farms (as determined based on the destinations of all batches that move from the market) with an equal probability. A graphic depicting how the network is altered by the inclusion of markets is shown in figure 3.3 and pseudo-code for the updated epidemic process is given as algorithm 2.

In the UK, livestock have to be removed four hours after the last market sale and consequently do not stay overnight at a livestock market [DEFRA, 2012]. Therefore, we assumed that cattle are moved on and off a market on the same day and that an infectious market becomes susceptible again the following day. If this assumption were to be relaxed the model could be altered by giving markets a longer

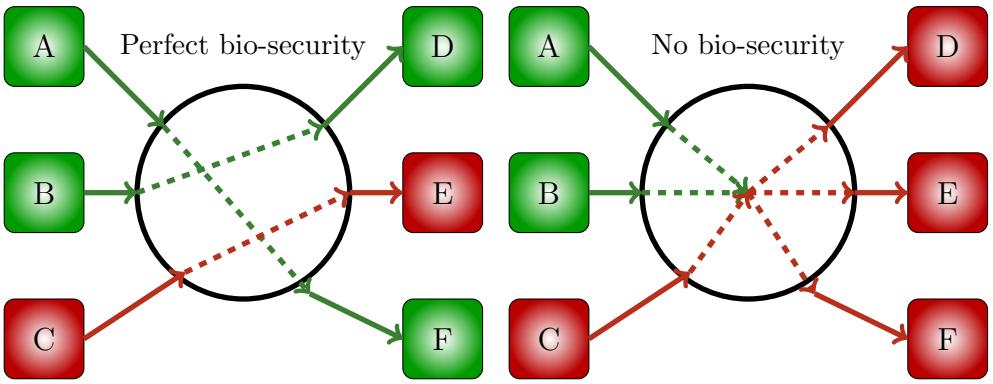

Figure 3.3: Two extremes of markets are included in this paper. In the first instance bio-security on a market is effectively assumed to be perfect, markets play no role in the transmission process. This is depicted in the left-hand figure where an infected movement (red) from farm C passes through the market but the only farm with a probability of being infected is its target farm, farm E. In the other extreme we assume that the market has effectively no bio-security and an infected movement from farm C has potential to infect the market. If it does so all farms receiving movements from the market have a non-zero probability of becoming infected.

Algorithm 2 Epidemic process including markets with an SIS model

1: **while** $\sum(I) > 0$ **do**
2: Infection loop for markets
3: Infection loop for farms
4: Recovery loop for farms
5: Set all markets to be susceptible
6: $t++$
7: **end while**

infectious period.

Simulations are carried out in the same way as detailed in the previous section with the one exception that we only run outbreaks for the length of one infectious period. The increased transmissibility from the inclusion of markets results in substantially larger epidemics and therefore one infectious cycle is sufficient to analyse the effects of the different sampling methods.

3.3 Results

3.3.1 Comparison of network statistics

As the percentage of nodes sampled decreases, the number of movements and nodes captured is observed to decrease for all sampling methods. As the network frag-

ments, the size of the GSCC decreases whilst the number of strongly connected components increases. Both the TNS and SBS sampling schemes out perform the RNS scheme in preserving robustness across all measured statistics as the percentage of nodes sampled decreases (figure 3.4). Within the giant component the mean degree and degree standard deviation remain robust with approximately 15-20% of the data for both TNS and SBS.

The complete network has assortativity coefficients close to zero, meaning there is not a tendency for similar nodes to connect to or avoid each other. This holds true within the GSCC for all node sampling schemes (figure 3.5). Similar behaviour is observed for the mean local clustering coefficient, which is small and does not change appreciably. The network diameter is 24 – TNS and SBS preserve this relatively small diameter well within the GSCC but for small sample sizes the diameter increases under the RNS scheme. Plots for assortativity, clustering and diameter are shown in 3.5.

Network statistics for the RMS sampling scheme as a function of movements sampled are given in figure B.1.

3.3.2 Comparison of epidemic predictions

In order to explore the epidemiological effects of the various sampling methods we compare each method with simulations run on the full network. The robustness of a sampling method is determined by whether the mean simulation for a method using a certain percentage of data lies within the 95% confidence intervals of the mean of simulations run in the full network. We focus on key epidemiological quantities such as final size, peak size and epidemic duration. Whilst it is informative to explore the effect of partial knowledge upon epidemic duration, for many diseases livestock movement bans will be implemented as soon as cases are reported. We therefore also look at predictions of the epidemic size after 6 and 12 weeks using the different sampling methods. We denote the threshold at which a scheme fails to be robust as $S_{\min}$, the minimum sampling threshold. Initially we set $\beta = 1$ and $T = 21$ days. Sensitivity to these parameter values is explored below (figures B.7 to B.8).

For outbreaks seeded in Cumbria and simulated on the full data set we obtain a final mean epidemic size of 185 farms, with a mean duration of 22 weeks and a mean peak size of 45 farms (figures 3.6 and B.3). The mean epidemic sizes after 6 and 12 weeks were 23 and 64 farms respectively. For all node sampling methods without rescaling, the epidemic size is under-predicted as the percentage of nodes sampled decreases (figure 3.6). After 6 weeks, $S_{\min}$= 3% for the TNS method, 20% for the SBS method and 80% for the RNS method. After 12 weeks, the percentage

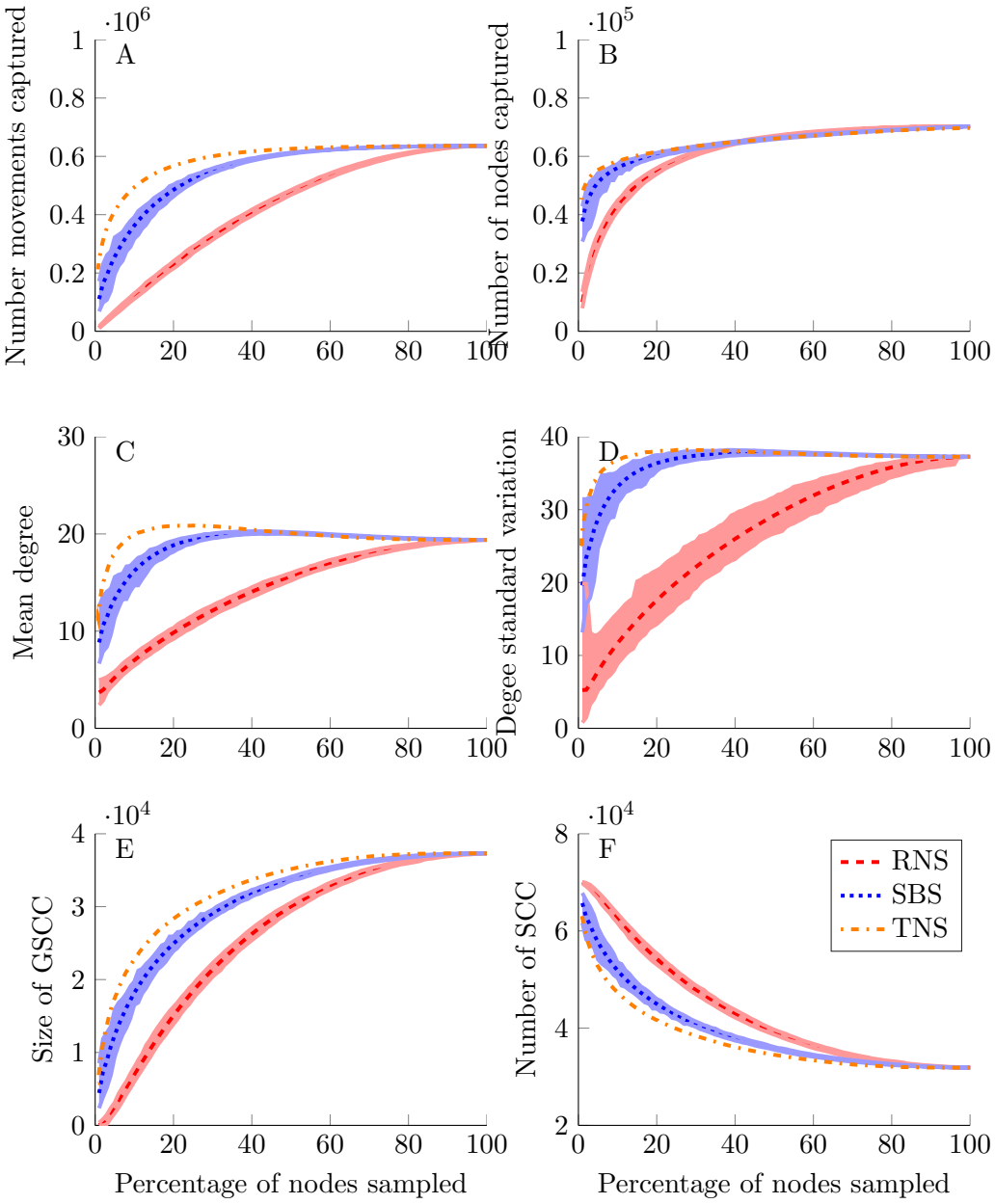

Figure 3.4: Graphs showing A - the number of movements captured, B - the number of nodes captured, C - the mean degree, D - the degree standard deviation, E - the size of the giant strongly connected component, and F - the number of strongly connected components for the RNS, SBS, and TNS methods as a function of the percentage of nodes sampled. These statistics are averaged over 1000 realisations of the network for RNS and SBS with shaded confidence intervals depicting the maximum and minimum value of each statistic.

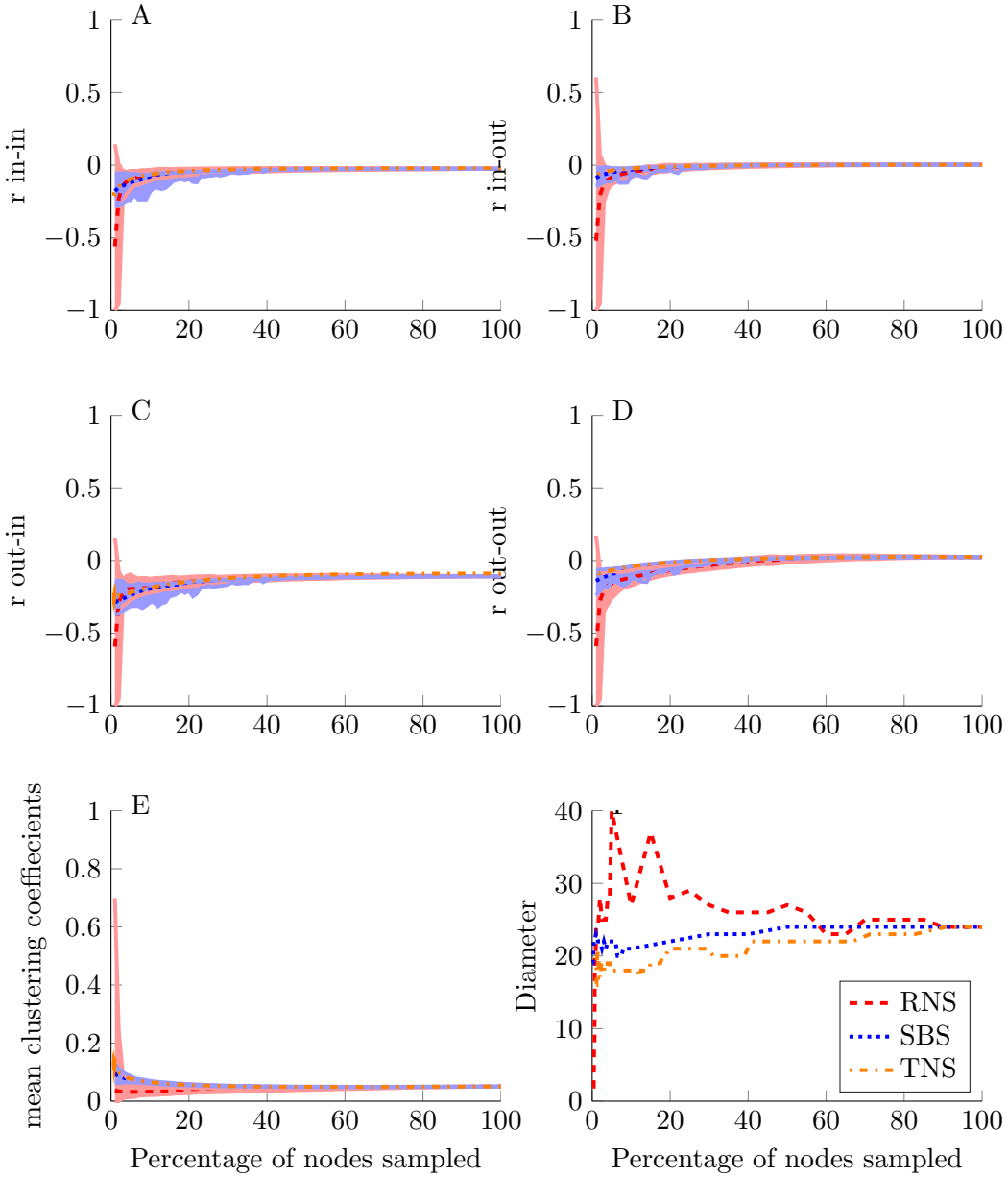

Figure 3.5: Graphs A, B, C and D show the directed degree-degree correlations for the three node sampling measures as the percentage of nodes sampled varies. The assortativity measures are defined using the Pearson correlation explained in the text above. The mean local clustering coefficient is plotted in E. The shaded regions on plots A to E indicate the maximum and minimum values obtained for these measured over 1000 realisations of the networks for RNS, SBS, and TNS. Plot F shows the network diameter for one realisation of the network for each of the sampling schemes.

of nodes that must be sampled increases to 9%, 30% and 90% for the TNS, SBS and RNS methods, respectively (figure 3.6-B). In order for these methods to accurately predict the full epidemic, 14%, 40% and 90% of the nodes must be sampled for the TNS, SBS and RNS methods respectively (figure 3.6-C). For the RMS method, $S_{\min}$= 30% for 6 weeks, 50% for 12 weeks and 80% for the whole epidemic.

Contour plots for epidemic size predictions for outbreaks seeded in Cumbria for each week of the outbreak (from week 1 to the end of the epidemic) are shown in supplementary figures B.3 to B.6. All methods provide accurate predictions of the size of the epidemic in the first few weeks. However, for longer durations, the TNS and the SBS methods provide the most robust predictions of epidemic size over time. Similar behaviour is observed for model predictions of epidemic duration and epidemic peak size (figures B.3 to B.6-B & B.6-C) – the TNS method is able to accurately capture these characteristics when only 15% of the nodes are sampled, compared with 30%, 80% and 90% for the SBS, RMS and RNS methods respectively.

The TNS method is consistently found to provide most accurate predictions of epidemic size, regardless of the county of disease introduction and disease parameters. In Devon, only 3% of the nodes require sampling for the TNS method to predict epidemic sizes at 6 weeks, compared with 10%, 20% and 80% for the SBS, RMS and RNS methods respectively, with similar effects seen at 12 weeks and for the full epidemic (figure B.9). Similar behaviour is observed in Aberdeenshire (figure B.11). The values for S_{min} for the full epidemic for all sampling methods for epidemics seeded in the three counties are summarised in table 3.1.

S_{min}	Cumbria	Devon	Aberdeen
RMS without markets	50%	40%	60%
SBS without markets	30%	20%	30%
TNS without markets	9% (50)	3% (80)	8% (60)
RMS with markets	20%	10%	10%
SBS with markets	50%	40%	50%
TNS with markets	30% (20)	25% (25)	30% (20)

Table 3.1: Summary of the minimum sampling threshold S_{min} after 12 weeks for random movement sampling (RMS), snowball node sampling (SBS) and targeted node sampling (TNS) without markets and S_{min} after one infectious period with markets for epidemics seeded in Cumbria, Devon and Aberdeen. Parameters were $\beta = 1$ and $T = 21$ days. For TNS the number of movements a farm needs in order top be sampled is given in parentheses.

As the transmission rate of the disease increases, epidemic sizes increase and a higher percentage of nodes are required for all sampling methods to make

accurate predictions. For example when $\beta = 2$, 25% of the nodes must be sampled using the TNS method and 50% for the SBS method to predict the overall epidemic size for outbreaks seeded in Cumbria, whilst for the RNS method almost all nodes must be sampled to capture epidemic behaviour. As the infectious period of the disease increases, a higher percentage of nodes needs to be sampled, but the effect of this is less pronounced than a variation in the transmission rate (supplementary figures chapter B). For diseases with a very high transmission rate, a much higher percentage of nodes must be sampled for all methods, even when the infectious period is short. Figures depicting the epidemic size for different parameter values can be found in the appendix (figures B.7 & B.8)

When we include within-market transmission into our model, we observe significantly larger epidemic sizes, with the mean epidemic size after one infectious period when $\beta = 1$ and $T = 21$ being 2,266 farms for outbreaks seeded in Cumbria. The TNS and SBS methods under predict epidemic sizes when less than 35% and 50% respectively of the nodes are sampled (figure 3.7). In contrast to the scenario where markets do not amplify transmission, the RMS method predicts epidemic sizes accurately even when only a very small number (approximately 20%) of movements are sampled. This suggests that, if a significant level of transmission is thought to occur within markets, then either TNS or RMS would be the preferred strategies if only limited resources were available. Similar results are observed for outbreaks seeded in Devon and Aberdeenshire.

3.3.3 Spatial spread

It is important to consider not only the size of the simulated epidemics but also how well the model captures the spatial spread of infection when partially sampled networks are used. When epidemics are seeded in Cumbria, almost all infected movements occur within Cumbria itself and to neighbouring counties (figure 3.8). An average of 8.9 farms become infected in Cumbria after 12 weeks with 13.9 in North Yorkshire and 4.2, 4.0, 4.0 and 8.0 in Durham, Lancashire, Dumfries and Galloway, and Aberdeen respectively. All other counties have epidemic sizes of fewer than 2 farms when the epidemic is seeded in Cumbria.

Using S_{min} for each of the sampling methods we find that: SBS captures the main epidemic hotspots well, but slightly overestimates epidemic sizes in these hotspots, RMS also performs well, but slightly underestimates epidemic sizes. The TNS method proves an accurate predictor of epidemic sizes in all the most infected counties, with 8.7 and 13.7 farms being infected on average after 12 weeks in Cumbria and North Yorkshire respectively (figures 3.8 & 3.9).

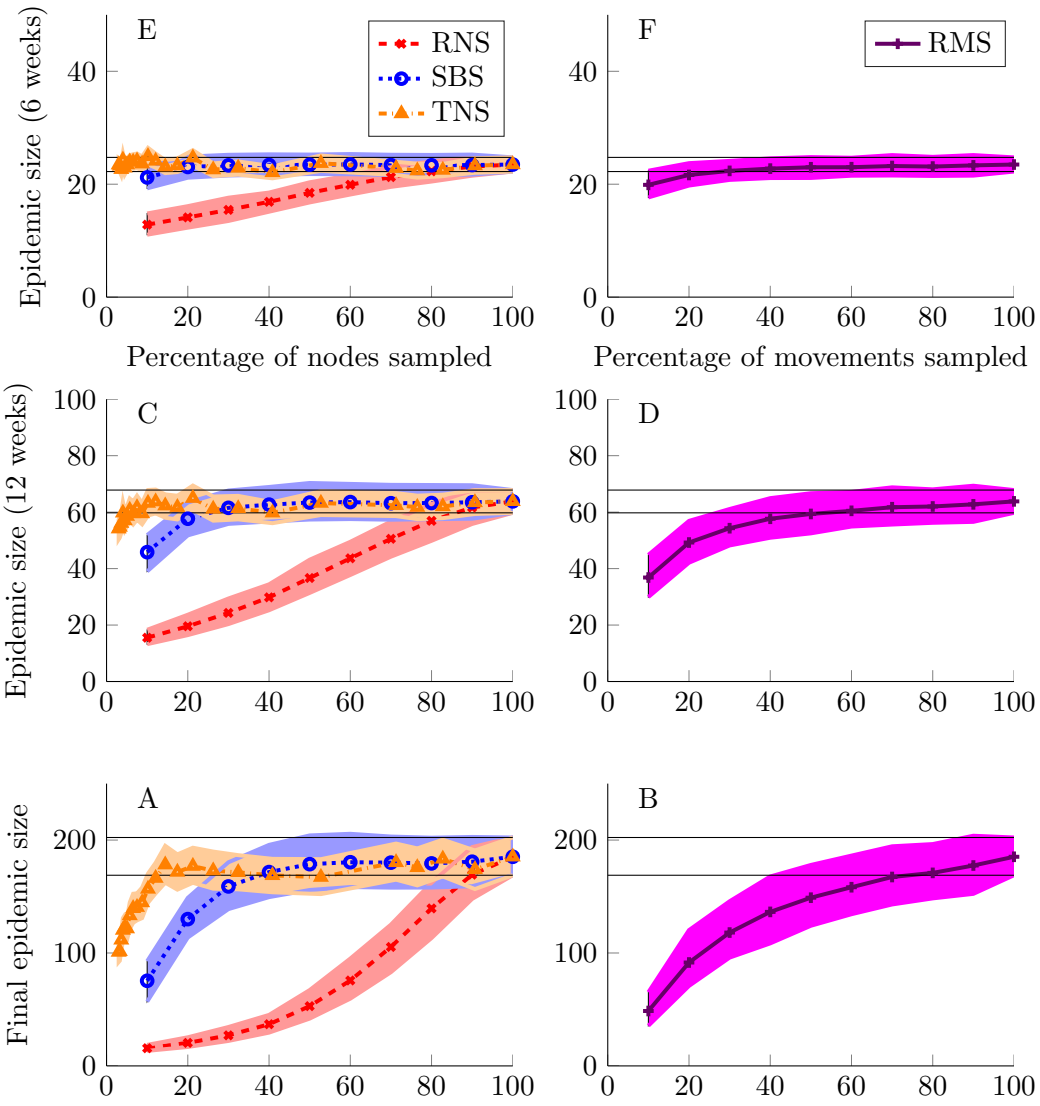

Figure 3.6: Graphs A, C, and E compare the epidemic size for outbreaks seeded in Cumbria on networks generated by RNS, SBS, and TNS as a function of nodes sampled with shaded 95% confidence intervals for A - 6 weeks, C - 12 weeks, and E - the full epidemic. The solid black lines represents the 95% confidence intervals on the average simulation for the original network. Graphs B, D, and F show the same results for the RMS method for B - 6 weeks, D - 12 weeks, and F - the full epidemic.

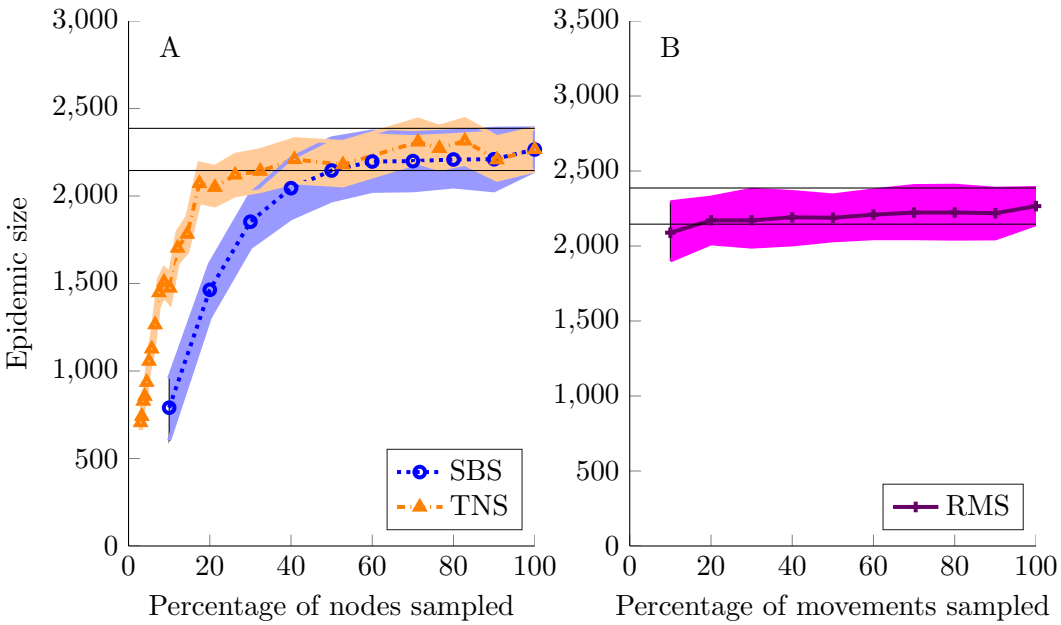

Figure 3.7: Mean epidemic size for outbreaks seeded in Cumbria, with within-market transmission incorporated into the model when $\beta = 1$ and $T = 21$, for A - the TNS (triangles) and the SBS (circles) methods, and B - the RMS (crosses) method. Shaded 95% confidence intervals are shown in each figure for each sampling method.

When markets are included the pattern of spatial spread is found to be similar to that without markets (figure 3.10). The three most highly infected counties on the full network are Cumbria, North Yorkshire and Aberdeenshire with mean epidemic sizes of 210, 258 and 238 respectively. When markets are included, we observe much larger epidemics in Devon, Somerset and North-East Wales. We also observe that each of the three sampling methods compares well with the original network at the S_{min} threshold.

Similar results are observed when outbreaks are seeded in Aberdeenshire and Devon transmission within markets results in outbreaks with a much large spatial extent than outbreaks in which markets do not play a role in transmission, figures B.13 - B.19.

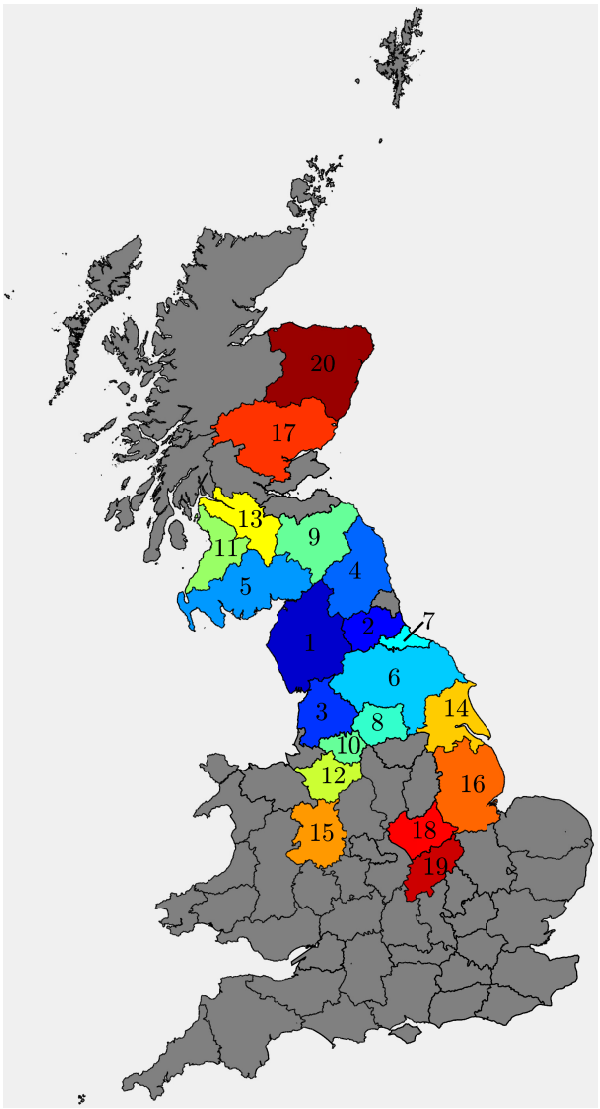

Figure 3.8: Map showing the 20 counties with the largest mean number of infected farms after 12 weeks when epidemics are seeded in Cumbria and markets are not explicitly included. The mean size for each county is given in figure 3.9.

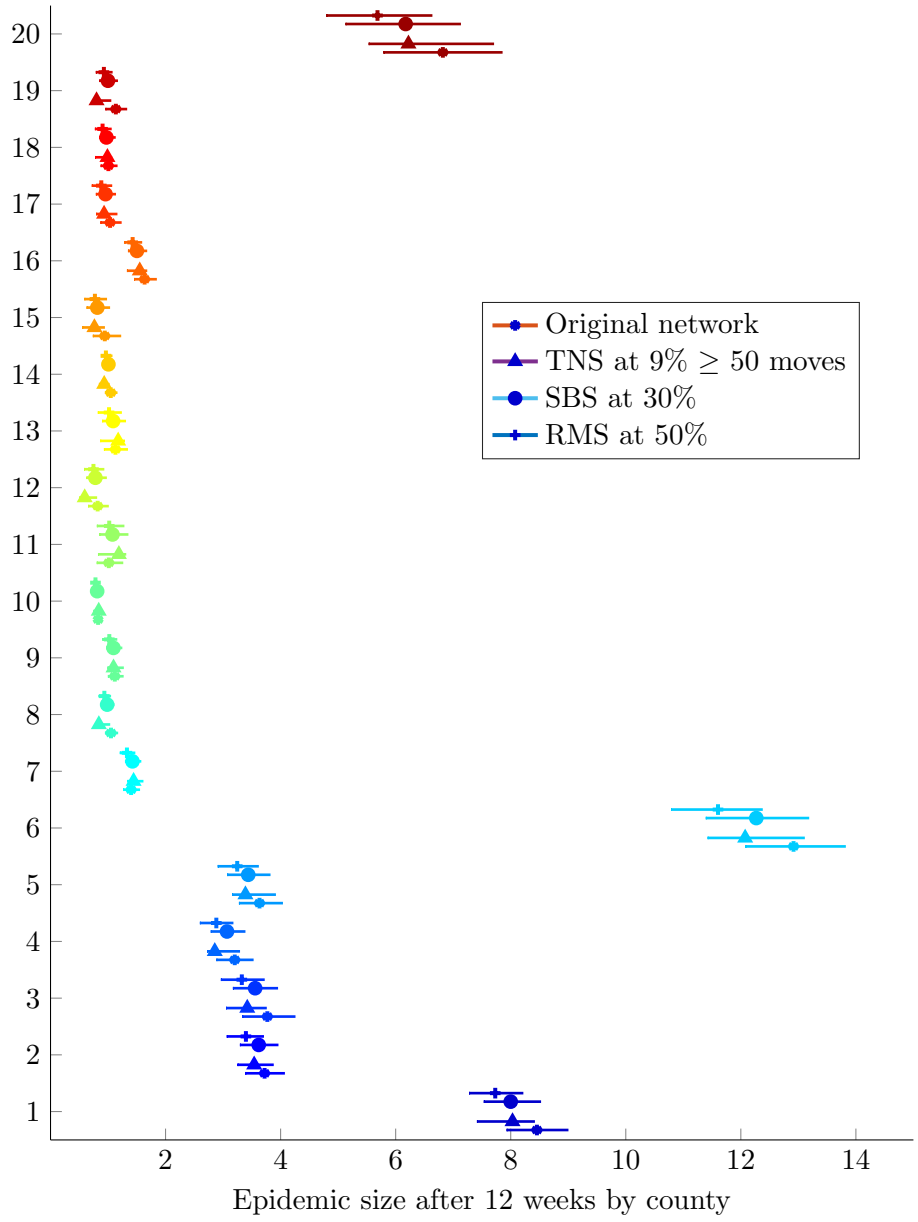

Figure 3.9: The average epidemic size for the original network (stars), random movement sampling (RMS) with 50% of sampled movements (crosses), snowball sampling with 30% of nodes (circles), and targeted node sampling (TNS), sampling nodes with more than 50 movements (triangles) for the 20 most infected counties when epidemics are seeded in Cumbria. Counties are ordered in terms of the proximity of their centroids from Cumbria with labels corresponding to the map in figure 3.8.

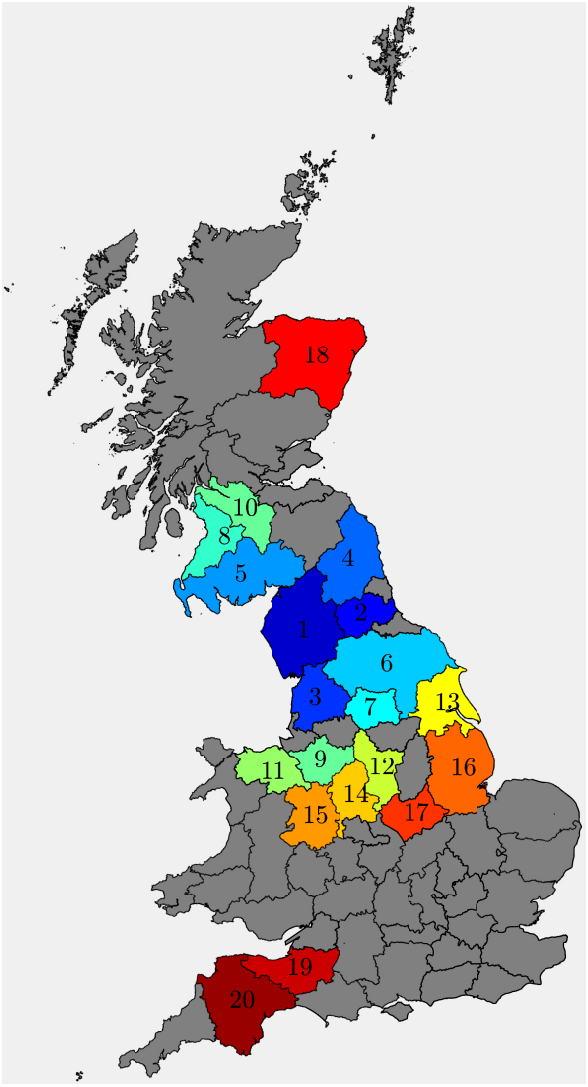

Figure 3.10: Map showing the 20 counties with the largest mean number of infected farms after one infectious period ((21 days) when epidemics are seeded in Cumbria and markets are explicitly included. The mean size for each county is given in figure 3.11.

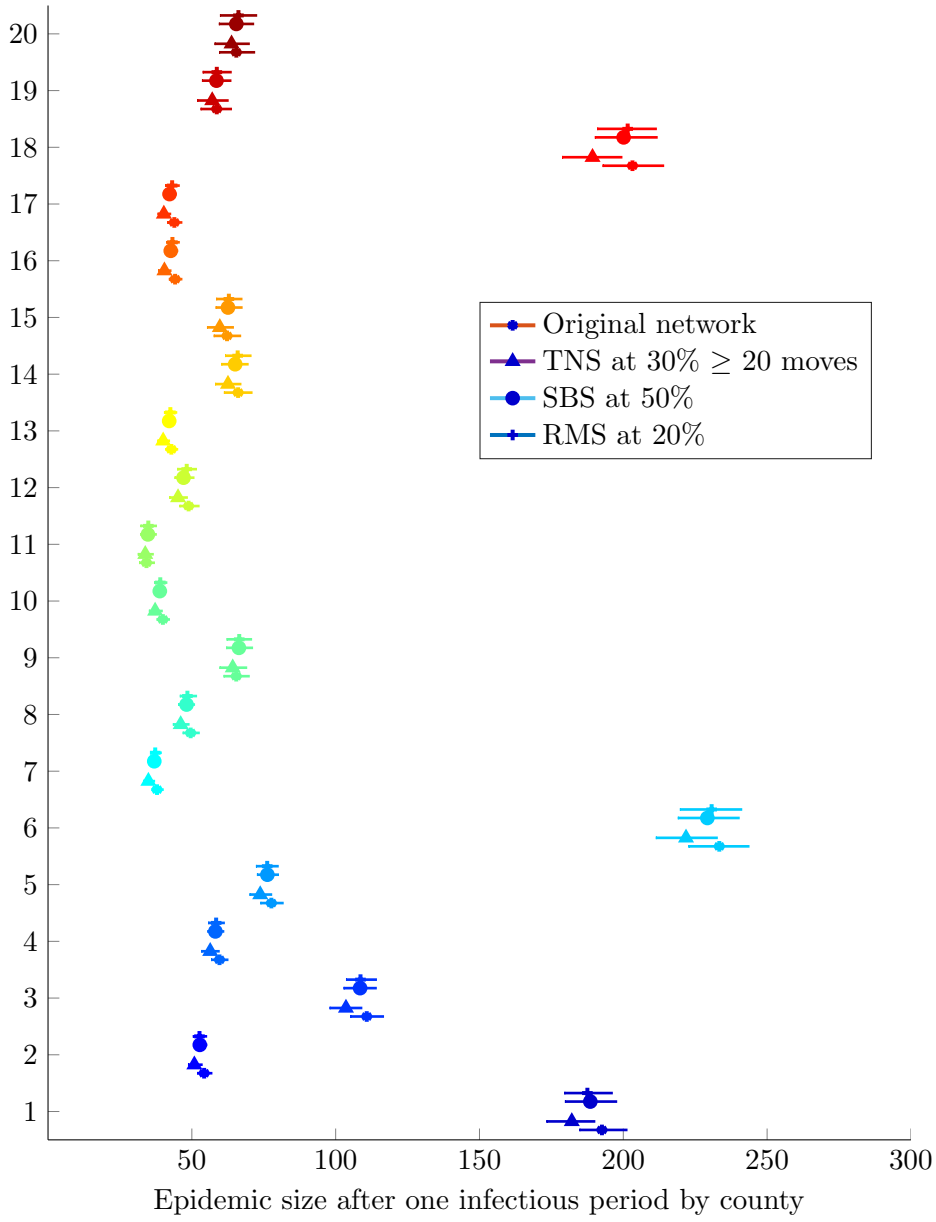

Figure 3.11: The average epidemic size for the original network (stars), random movement sampling (RMS) with 20% of sampled movements (crosses), snowball sampling with 50% of nodes (circles), and targeted node sampling (TNS), sampling nodes with more than 20 movements (triangles) for the 20 most infected counties when epidemics are seeded in Cumbria. Counties are ordered in terms of the proximity of their centroids from Cumbria with labels corresponding to the map in figure 3.10.

3.4 Discussion

In order for models to be used to predict the potential for disease spread in livestock, there is a reliance upon accurate data regarding farm locations and movements of livestock between farms. Significant work has been done in the UK to predict the potential for disease spread through the livestock network e.g [Brooks-Pollock and Keeling, 2009; Green et al., 2006; Ortiz-Pelaez et al., 2006] owing to the existence of the cattle tracing scheme and the animal movement license scheme. In many other countries around the world, the lack of such databases means that it is impossible to develop a model that utilizes precise movement data and an alternate approach must be used. In such countries, it may be impossible to ever record all movement data either owing to the sheer size of the industry (in countries such as the USA) or owing to the cost associated with implementing an animal license scheme. However, a more limited data collection scheme may be possible, whereby movements are recorded for a subset of the livestock movement network.

A simple way to collect a subset of livestock movement data would be to randomly sample all movements from a given set of random farms (i.e. using the RNS method). This method proved ineffective at reproducing the mean epidemics seen on the complete network. An alternate strategy to collect movement data would be to randomly sample movements from any farm (i.e. the RMS method). In a practical sense, this would be a much more difficult strategy to implement, requiring individual farmers to keep a record of livestock moving from their farm a given percentage of the time. This method is found to be more effective than the RNS method, particularly in the case when within-market transmission occurs. In that case, only 10-20% of movements are required in order to accurately predict epidemic sizes. For lower percentages, the model predicts smaller epidemics than observed using the true network data and in that case, suggested intervention strategies may not be sufficient to control outbreaks. It may, however, be possible to make accurate predictions with a lower percentage of movement data when incorporating a Bayesian kernel approach to scale up a partially observed network [Lindström et al., 2013]. An alternative approach may be to adopt targeted movement sampling where movements would be recorded based on some criterion. For example, particularly frequent movements between pairs of nodes could be recorded, or shipments involving a large number of animals. Both of these options were investigated but neither proved to be particularly successful at reconstructing an accurate realization of the original network.

If only limited resources are available for data collection, it may be more

efficient to record movements only from the most highly connected farms (the TNS method) or to use snowball sampling (the SBS method). The TNS method proves significantly more effective than both the RMS and RNS methods when markets do not contribute to transmission – less than 20% of all farms would need to be surveyed in order to predict epidemic sizes to within 90% confidence in the UK for outbreaks seeded in Cumbria, Devon and Aberdeenshire. The model also gives a very good approximation of the spatial spread of the disease, the size of the epidemic peak and the epidemic duration. When markets contribute towards disease amplification, the TNS method requires that around 30% of all be surveyed in order to accurately predict epidemic sizes. The SBS method is found to perform less effectively than the TNS method, as this strategy rapidly identifies the most highly connected nodes that are likely to contribute most significantly to disease transmission. However, the SBS method may be more practical to implement as it does not require prior knowledge of the relative connectivity of the farms in the network.

The TNS and SBS methods have worked favourably in the livestock network described here. Whilst one must take care when making inference from a sub-network to a full network [Stumpf et al., 2005], it would be of great interest to the broader study of disease spread on partially observed networks to test these strategies further on livestock networks such as those available in other European countries [Bajardi et al., 2012; Nöremark et al., 2009; Rautureau et al., 2012]. The results of this work provide evidence of the viability of using partially sampled data to predict disease spread in livestock [Buhnerkempe et al., 2013] and humans [Danon et al., 2009; González et al., 2008; Pindolia et al., 2013; Wesolowski et al., 2013] and will inform data collection strategies in situations where complete knowledge of the network is impossible (e.g. wildlife movements [Odden et al., 2014; Robinson et al., 2012]).

The role played by markets in disease transmission may have a significant effect upon the predictability of the sampling methods. When markets do not contribute to disease transmission, only a very small percentage of nodes needs to be sampled using the TNS method. However, when we make the assumption that all batches on a market are well-mixed, a much large proportion of the nodes must be sampled. We also find that in this case, the RMS method requires sampling of a much smaller percentage of movements than the non-market scenario. This is unsurprising – markets represent very highly connected nodes in the network and therefore when they are explicitly included in the model, an RMS approach will preferentially sample movements to and from these highly connected nodes. The model currently assumes that livestock do not stay overnight on markets, in

line with Defra policy, and hence any infectious markets would become susceptible the following day. Should this not be the case, the role of markets in disease transmission may be slightly altered. Therefore, our results suggest that a more thorough, disease-specific, analysis of the precise role of markets in disease transmission would be required in the future in order to determine context-specific optimal sampling strategies. However, our sensitivity analysis shows that TNS is the preferred sampling strategy for all studied transmission rates and infectious periods.

The model presented in this chapter uses a weighted static network to simulate the risk of transmission between livestock farms. Weighted static networks are regularly used in livestock disease models and previous work indicates that they provide good prediction of mean epidemic sizes, though may potentially underestimate variability when compared with results on dynamic networks [Vernon and Keeling, 2009]. The advantage with a weighted static network approach is that it is possible to determine the epidemic impact independent of time of year. However, there is clear seasonality observed in the cattle movement network [Kao et al., 2007] and it is therefore possible that a weighted static network could result in an under or over prediction of epidemic size. Our sensitivity analysis suggests that the preferred sampling strategies are robust, although the proportion of nodes that need to be sampled may vary dependent upon time of year. Future studies will focus upon constructing a dynamic network and testing network sampling schemes and temporal sampling schemes (whereby sampling is targeted based on time of year) on their ability to predict epidemic behaviour.

The results indicate that for a fast spreading disease such as FMD, sampling a small proportion of the network is sufficient. This relies on the assumption that infected movements result in all livestock on the destination farm becoming rapidly infected. This is not the case for all livestock diseases. For example, animals infected with bTB can remain asymptomatic carriers for several months [Brooks-Pollock et al., 2014] before becoming infectious. The model framework described here would not be appropriate for a disease of this nature and further work will focus upon the development of optimal sampling strategies for slow spreading diseases such as bTB where it may be crucial to track movements of individual cattle.

Our results suggest that for countries with similar farming practices, it may not be necessary to collect data on all livestock farms, but only those that contribute most significantly to the livestock trade network. Of course, this creates something of a conundrum – in order to sample the most highly connected nodes and thus accurately represent epidemic risk on an unknown network, one needs to know which farms have the most movements. One solution to this would be for all farmers to

be required to record the number of movements they make in a given year. These summary statistics could then be used to determine which farms should be sampled for the following year. In the UK at least, analysis of the movement network for multiple years suggests that those farms that have a high number of movements in a given year are more likely to have a high number of movements in the following year. This method may therefore be utilized in countries where livestock movement data are not currently available in order to inform epidemic models and predict the potential for disease spread owing to animal movements in the early stages of a disease outbreak.

Chapter 4

Community detection and epidemic predictions

> "I'm very glad you asked me that, Mrs Rawlinson. The term 'holistic' refers to my conviction that what we are concerned with here is the fundamental interconnectedness of all things. I do not concern myself with such petty things as fingerprint powder, tell-tale pieces of pocket fluff and inane footprints. I see the solution to each problem as being detectable in the pattern and web of the whole. The connections between causes and effects are often much more subtle and complex than we with our rough and ready understanding of the physical world might naturally suppose, Mrs Rawlinson.
> "Let me give you an example. If you go to an acupuncturist with toothache he sticks a needle instead into your thigh. Do you know why he does that, Mrs Rawlinson?
> "No, neither do I, Mrs Rawlinson, but we intend to find out. A pleasure talking to you, Mrs Rawlinson. Goodbye."
>
> Douglas Adams, Dirk Gently's Holistic Detective Agency [1987]

The networks that we have created from farms and cattle movements can be analysed in many ways from looking at distributions to computing various metrics and measures. Of particular interest to epidemic spread is the degree of clustering in the network. High clustering values can lower R_0 and lead to smaller epidemics than one would find in a network with a similar degree distribution but a lower clustering coefficient [House and Keeling, 2011; Keeling, 2005b; Szendroi and Csányi, 2004] as the epidemic becomes contained by a lack of local susceptible nodes; however later in the epidemic high clustering can lead to pockets of susceptible nodes.

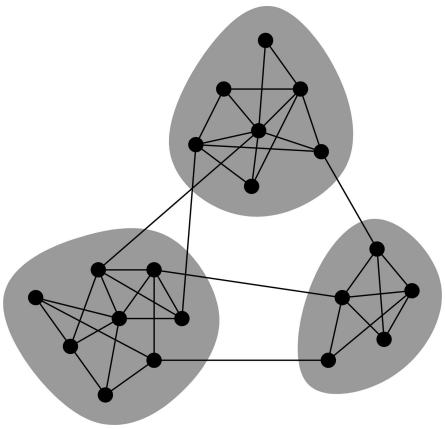

Figure 4.1: A network divided into three communities, edges are denser within the communities than they are between communities. This figure is reproduced directly from Newman [2006].

Clustering is a local effect; a more complex analysis is to consider the community structure of the network. The concept of a community is understood in social context as a group of people that live in the same area (neighbourhood, city, country) or share common interests (religion, race, computer gaming etc..). In this social context we can think of every individual as a node in our network and a link can be made between nodes if they have common interests or live near one another. Once this network is formed it is interesting to see if a community structure exists in this network, are there cliques, or groups that form? A division of nodes into communities would see groups of densely connected nodes with sparser connections between groups (figure 4.1).

In this chapter we define what we mean by a community and community structure, discuss an algorithm for community detection and implement this algorithm on the UK and Turkey cattle networks. We then explore the predictions that community structure can have on the spread of epidemics by running epidemic simulations on the cattle network of the UK and comparing the results with the predictions offered to us by the community structure of the network.

4.1 Theory

The discovery of groups in a network generally follows two principal lines of research, graph-partitioning which is typically looked at by computer scientists [Buluç et al., 2013], and community detection [Fortunato, 2010] which has applications in social [Arenas et al., 2004] and biological networks [Guimerà and Amaral, 2005]. While the

two problems are similar in nature, they both search for divisions of a network into groups, there is a fundamental difference. Graph partitioning typically deals with dividing a set of tasks, say, amongst a number of processors in a parallel computer, the number of processors and the number of tasks each processor can perform is typically known so the problem is to the divide the tasks amongst the processors such that the number of links between the processors is minimal. In community detection, however it is not known *a priori* how many groups are being searched for, how big they should be, or even if it is appropriate to partition the network in the first place.

4.1.1 Definition of modularity

Let us begin with a symmetric network with N nodes described by an adjacency matrix

$$A_{ij} = A_{ji} = \begin{cases} 1, & \text{if a link between } i \text{ and } j \text{ exists} \\ 0, & \text{otherwise.} \end{cases} \tag{4.1}$$

with $m = \sum_{ij} A_{ij}/2$ edges. This network has been divided into a partition $\mathcal{P}$, consisting of groups which we call communities. We wish to evaluate if the partition presented to us displays a strong community structure.

First we consider what a 'strong' partition of a network into a community would be. To do this we define the modularity Q of a partition $\mathcal{P}$. First proposed by Newman [2004], modularity measures if there are more connections between communities than one would expect to see by chance alone. That is, we assume that a random graph does not display a strong community structure and compare the given community structure with one where the communities are the same size but the edges are placed at random,

$$\begin{aligned} Q = (&\text{fraction of edges within communities}) - \\ &(\text{expected fraction of such edges}) \end{aligned} \tag{4.2}$$

This expected fraction depends on the chosen null model [Fortunato, 2010]. The null model would be a copy of the graph that keeps some structural properties from the original graph, such as degree distributions, but does not have a community structure.

The contributions to the modularity from each pair of nodes within a community C is summed. A pair of nodes within a community gives a greater contribution to Q if the strength of the link between the pair is greater than one would

expect from a random graph drawn from the null model. If nodes are in the same community but no link exists between the two nodes their contribution to Q would be negative. A sum is then taken over all of the communities in $\mathcal{P}$. Recasting equation 4.2 as a mathematical expression

$$Q = \frac{1}{2m} \sum_{C \in \mathcal{P}} \sum_{i,j \in C} [A_{ij} - P_{ij}] \tag{4.3}$$

where P_{ij} is the null model which represents the expected weight of an edge between nodes i and j over an ensemble of random networks with certain constraints. Modularity is effectively counting the number of links within in a community compared to what one would expect to find in the null model.

Here, as is customary, we have normalised the modularity by dividing by the total strength of the network $2m$ so that it ranges from -1 to 1. Higher values of Q indicate a stronger community structure but the maximum value of Q is dependent on both the network under consideration and the choice of null-model. The closer the null model is to the network under investigation the lower the value of Q found for that network. Because of this when comparing the community structure found between different networks and different null-models we can not definitively say if one network exhibits a stronger community structure than another, the results of modularity optimization must be interpreted in a more descriptive fashion.

4.1.2 Choice of null model

The choice of null model $\mathbf{P}$ will determine the value of Q and will therefore have consequences for any partition found via modularity optimization. For the moment we will deal only will un-weighted (edges are binary), symmetric (undirected) network described by equation 4.1.

A very basic null model would come from the constraint that we require the null model to have the same number of edges as the original graph with edges placed randomly between nodes, describing a Bernoulli random graph and yielding a constant null model term in equation 4.3

$$P_{ij}^{\text{Bernoulli}} = \frac{2m}{N(N-1)}. \tag{4.4}$$

The Bernoulli random graph would however, be a poor choice of null-model as it has a Poission degree distribution which is very different from the degree distributions found in real world networks [Albert and Barabási, 2002; Arenas et al., 2004].

We wish to choose a null model that preserves the degree distribution of the

original network taking into account the degree of each node $k_i = \sum_j k_{ij}$. The most common choice of null model is the Newman-Girvan model [Newman and Girvan, 2004]

$$P_{ij}^{\mathrm{NG}} = \frac{k_i k_j}{2m} \tag{4.5}$$

which is similar to the configuration model of networks [Newman, 2010]. This choice is quite natural as it takes into account the heterogeneity of the network and obviously penalises unsurprising connections; if two nodes both have high degrees then the probability of the two being connected is also high so their contribution to the modularity will be low compared to that of two nodes with low degrees. The modularity function defined from this choice of null model we will denote as Q^{NG}.

The Newman-Girvan null-model has been applied to a diverse range of networks such as; character interactions in Alexander Dumas famous *Les Miserables* [Newman and Girvan, 2004] where the 11 communities found represented subplots of the novel, product recommendations on Amazon [Clauset et al., 2004] which allowed for classification of products into categories, co-authorship links between academics [Newman and Girvan, 2004], and mobile phone networks [Expert et al., 2011].

4.1.3 Edge direction and weight

Extending the definition of Q^{NG} to directed weighted networks is straight forward [Newman, 2004]. We first redefine the adjacency matrix as

$$A_{ij} = \begin{cases} w_{ij}, & \text{if a link goes from } j \text{ to } i \\ 0, & \text{otherwise.} \end{cases} \tag{4.6}$$

where w_{ij} is weight of the link from node i to node j, we note that in general $A_{ij} \neq A_{ji}$. We now define the weighted in and out degrees for each node as $w_i^{\mathrm{in}} = \sum_j A_{ij}$ and $w_i^{\mathrm{out}} = \sum_j A_{ji}$. The total weight of the network is now $w = \sum_i w_i^{\mathrm{out}} = \sum_i w_i^{\mathrm{in}}$. If all the weights equal one then the weighted degrees are equivalent to the in and out degrees for each node. The definition of Q^{NG} becomes

$$Q^{\mathrm{NG}} = \frac{1}{w} \sum_{C \in \mathcal{P}} \sum_{i,j \in C} \left[A_{ij} - \frac{k_i^{\mathrm{in}} k_j^{\mathrm{out}}}{w} \right] \tag{4.7}$$

or equivalently

$$Q^{\mathrm{NG}} = \frac{1}{w} \sum_{i,j} \left[A_{ij} - \frac{k_i^{\mathrm{in}} k_j^{\mathrm{out}}}{w} \right] \delta(c_i, c_j) \tag{4.8}$$

where c_i is the label of the community to which vertex i belongs, (note that the factor of two is no longer in the denominator).

4.1.4 Coupling through time

The cattle networks of the UK and Turkey are dynamic in nature, movements are recorded on a daily basis, we can therefore construct a new network for every day of the year. These networks can be aggregated to form coarser networks at different temporal scales (weekly, monthly, seasonally, etc..) and a discussion of what resolution is appropriate is dependent on the question under investigation.

For now let us assume we have taken a year's worth of data, created twelve monthly adjacency matrices and ran a community detection algorithm on each to find partitions $\mathcal{P}_i$, $i \in \{1, .., 12\}$. Each $\mathcal{P}_i$ may have a different number of communities and thus there is no way to label a community in one partition such that it may be compared with a community in a different partition. Aspects of the partitions may be compared (i.e. number of communities, size of communities, locations, areas) but as things stand, the evolution of communities through time can not be explored. To compare communities in temporal partitions we need a coupling of the different networks through time. A method to do this was developed by Mucha et al. [2010] and it is this method we will use to couple networks through time.

Given a multi-slice network, i.e. a network that can be represented by S multiple slices each with its own adjacency matrix A_{ijs} defined with the same set of nodes, where the index s represents a slice. We couple nodes to themselves in different slices of the network with a coupling C_{jrs} that couples node j to itself between slices r and s. In general a node could be coupled with itself in every slice but as we are focusing on time-dependent networks we can order the slices through time and only create couplings between neighbouring slices (figure 4.2). The strength of the coupling could vary for different nodes and across different slices but we have no intuitive reason to do this so will keep the coupling strength fixed at ω; thus for every node j,

$$C_{jrs} = \begin{cases} \omega, & \text{if } s = r \pm 1 \\ 0, & \text{otherwise.} \end{cases} \tag{4.9}$$

in this way coupling is also symmetric $C_{jrs} = C_{jsr}$. This introduces some new notation which we need to derive the null-model for multi-slice networks, assuming an undirected network ($A_{ijs} = A_{jis}$), each nodes strength in an individual slice is now $k_{js} = \sum_i A_{ijs}$ and its strength across slices is $c_{js} = \sum_r C_{jrs}$ giving a multi-slice

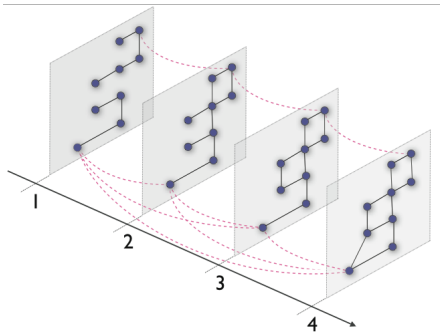

Figure 4.2: Schematic of a multi-slice network. Four slices s=1, 2, 3, 4 represented by adjacencies A_{ijs} encode intra-slice connections (solid lines). Inter-slice connections (dashed lines) are encoded by C_{jrs},specifying the coupling of node j to itself between slices r and s. For clarity, inter-slice couplings are shown for only two nodes and depict two different types of couplings: (i) coupling between neighbouring slices, appropriate for ordered slices; and (ii) all-to-all inter-slice coupling, appropriate for categorical slices. Reproduced from Mucha et al. [2010].

strength of $\kappa_{js} = k_{js} + c_{js}$. The multi-slice modularity is

$$Q_{\text{multi-slice}} = \frac{1}{2\mu} \sum_{ijsr} \left[\left(A_{ijs} - \frac{k_{is} k_{js}}{2m_s} \right) \delta_{sr} + \delta_{ij} C_{jsr} \right] \delta(g_{is}, g_{jr}) \tag{4.10}$$

where $2\mu = \sum_{jr} \kappa_{jr}$ and the delta function keeps contributions from the network edges and the artificial coupling separate. The multi-slice modularity as given here is clearly analogous to the Newman-Girvan modularity for a single-slice network, we extend this to a general multi-slice modularity for a directed network

$$Q^{\text{directed}}_{\text{multi-slice}} = \frac{1}{\mu} \sum_{ijsr} \left[\left(A_{ijs} - P_{ijs} \right) \delta_{sr} + \delta_{ij} C_{jsr} \right] \delta(g_{is}, g_{jr}) \tag{4.11}$$

where $\mu = \sum_{jr} \kappa_{jr}$.

4.1.5 Finding the partition

As alluded to above the purpose of defining modularity is to aid us in finding the partition of the graph that gives us the 'best' community structure. The 'best' or optimal community structure for the network will be one that maximises the modularity function. Numerous methods are available to do this. From spectral methods [Leicht and Newman, 2008; Newman, 2006], to greedy algorithms and simulated annealing [Fortunato, 2010].

Here we will focus on the method known as the Louvain method which was

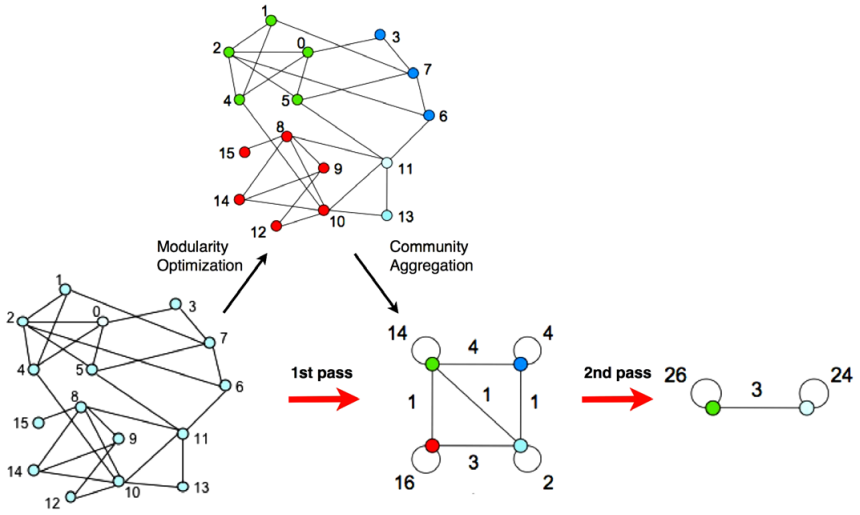

Figure 4.3: Graphical depiction of the Louvain method at work. The first phase consists of modularity optimisation while the second involves community aggregation. The combination of the two phases form a 'pass' and passes continue until modularity is maximised. Reproduced from Blondel et al. [2008].

first proposed by Blondel et al. [2008] and named for the Univeristé Catholique de Louvain where the authors were based when they devised this method. This method was chosen for its speed and also as the method was adapted for multi-slice networks by Mucha et al. [2010] and it is their MatLab code that we use through-out this study [Jutla et al., 2014].

The method consists of two phases that are repeated iteratively until no gains to the modularity can be found, see figure 4.3.

4.1.6 The first phase

Given N nodes each node is initially assigned to its own community, the initial partition therefore has as many communities as it does nodes. The first phase begins by considering each node i and its neighbours j. The respective changes in modularity ΔQ found by moving node i into each node j's community are calculated. Node i is then moved into the community which results in the maximum ΔQ, provided this change is positive. If $\Delta Q_{\max} \leq 0$ then node i does not move. This process is applied repeatedly and sequentially until no positive changes in the modularity can be found. Nodes may be considered multiple times as some may not see any gain in moving until others have moved first. The first phase ends when a local maxima of Q has been reached.

For this first phase we need to be able to quickly calculate the ΔQ in two

cases; i) when an isolated node is moved into a new community and ii) when a node is moved from one community to another. We will calculate this for the general form of the modularity given by equation 4.3. Assuming we have K communities we can write the modularity as a sum of the modularity contributions from each community

$$Q = Q_1 + Q_2 + Q_3 + \cdots + Q_{K-1} + Q_K \tag{4.12}$$

now let us assume that community c_K consists of only node k and we wish to calculate the change in modularity found by moving node k to community c_1 changing community c_1's modularity to Q_1'. The new modularity is now given by

$$Q' = Q_1' + Q_2 + Q_3 + \cdots + Q_{K-1} \tag{4.13}$$

and the change in modularity is

$$\Delta Q_{k\to c_1} = Q' - Q = Q_1' - Q_1 - Q_K. \tag{4.14}$$

The modularities of interest are

$$Q_1 = \frac{1}{2m}\sum_{i,j\in c_1}[A_{ij} - P_{ij}] \tag{4.15}$$

$$Q_1' = \frac{1}{2m}\sum_{i,j\in (c_1\cup k)}[A_{ij} - P_{ij}] \tag{4.16}$$

$$Q_K = \frac{1}{2m}A_{kk} - P_{kk} \tag{4.17}$$

which gives

$$\Delta Q_{k\to c_1} = \frac{1}{2m}\sum_{i\in c_1}[A_{ik} + A_{ki} - (P_{ik} + P_{ki})] \tag{4.18}$$

which reduces to

$$\Delta Q_{k\to c_1} = \frac{1}{m}\sum_{i\in c_1}[A_{ik} - P_{ik}] \tag{4.19}$$

assuming a symmetric null model. For a directed network $\Delta Q_{k\to c_1}$ drops the factor of a half from equation 4.18.

4.1.7 The second phase

In the second phase we make a new network out of the communities found during the first phase. Each community becomes a node in the new network with its edges to nodes in different communities combined together, edges within the community

are also combined to make self-loops. Once this has been completed the first phase can then be performed again and then iterated. Preforming the two phases is known as a 'pass'. Passes continue (figure 4.3) until there are no more changes that can increase the modularity, thus a maximum modularity is obtained.

4.1.8 Observations on the Louvain method

1. Extending the Louvain method to multi-slice networks is straightforward. The adjacency matrices from each slice can be merged into a super $(N\times S)\times(N\times S)$ adjacency matrix where each individual adjacency matrix is placed in blocks on the super-adjacency matrix such that node i rather than being represented by the i^{th} row and column in the original matrices will now be represented by the $(s-1)N+i$, $s=1,\ldots,S$ rows and columns in the super-adjacency matrix. We are only concerned with ordered next-slice coupling which is encoded symmetrically in the super-adjacency matrix as

$$A_{[(s-1)N+i][sN+i]} = A_{[sN+i][(s-1)N+i]} = \omega. \tag{4.20}$$

2. The first phase of each pass of the algorithm was to end when the local maxima of the modularity was attained. In practice the overall runtime of the process is decreased by introducing a threshold such that if any relatives gains in modularity are less than that threshold the first phase ceases.

3. The algorithm naturally creates a hierarchy as it finds communities. Each node starts in a community of size one before possibly being merged into a larger community, these communities are in turn then merged until no gain in modularity can be gained. By outputting the communities found after each phase it is possible to examine the overall community at different hierarchical levels.

4. It is important to note that the order the nodes are considered can alter the partition found by the algorithm and as such the algorithm should be run multiple times and the partitions examined before deciding on a definitive partition for the network.

4.1.9 Comparing partitions

As noted above the Louvain method for detecting community structure is not deterministic. To account for this we run the algorithm many times on the same network and save each partition $\mathcal{P}_i$. These partitions can be compared in several

ways; most commonly by a pair counting approach where the number of pairs of nodes appearing in the same community is calculated [Karrer et al., 2008; Traud et al., 2011] or by an information theoretic approach [Danon and Díaz-Guilera, 2005; Karrer et al., 2008; Meil, 2007]. We opt for the information theory approach given by Meil [2007] and advocated by Karrer et al. [2008]. This method is known as the variation in information VI[i] and is outlined below.

Given a partition $\mathcal{P}$ of N nodes divided into K communities $C_1, \ldots, C_K$ with sizes $n_1, \ldots, n_K$ if we were to pick a node at random the probability of that node being in community C_k is

$$P(k) = \frac{n_k}{N} \tag{4.21}$$

defining a discrete random variable which takes K values. The uncertainty associated with our partition $\mathcal{P}$ is given by the entropy [Shannon, 1948] of this random variable

$$H(\mathcal{P}) = -\sum_{k=1}^{K} P(k) \log P(K). \tag{4.22}$$

Mutual information between two partitions $\mathcal{P}$ and $\mathcal{P}'$ is the information that one partition has about the other and is given by

$$I(\mathcal{P}, \mathcal{P}') = \sum_{k=1}^{K} \sum_{k'=1}^{K'} P(k, k') \log \frac{P(k, k')}{P(k)P(k')} \tag{4.23}$$

the variation in information is then defined as

$$\mathrm{VI}(\mathcal{P}, \mathcal{P}') = \big[H(\mathcal{P}) - I(\mathcal{P}, \mathcal{P}')\big] + \big[H(\mathcal{P}') - I(\mathcal{P}, \mathcal{P}')\big] \tag{4.24}$$

and can be interpreted as the sum of the information needed to describe $\mathcal{P}$ given $\mathcal{P}'$ and the information needed to describe $\mathcal{P}'$ given $\mathcal{P}$. By dividing by $\log n$ we can normalise VI (we will always normalize VI); values close to one indicate the partitions are quite different while values close to zero indicate that the two partitions are similar.

When comparing a selection of partitions N_P at once we compute the normalized variation in information for $\mathcal{P}_i$ with all other partitions $\mathcal{P}_j$ and then average this to obtain the average normalized variation in information for partition $\mathcal{P}_i$

$$\langle \mathrm{VI}(\mathcal{P}_i) \rangle = \frac{1}{N_P} \sum_j \mathrm{VI}(\mathcal{P}_i, \mathcal{P}_j) \tag{4.25}$$

[i] As opposed to other approaches variation in information is a true metric in that it obeys the triangle inequality.

which can be compared across all partitions.

Taking the average VI found for each partition compared to every other partition gives us a method for choosing the most representative partition found, this will be the partition with the smallest average VI.

4.2 Community detection in the UK

We now seek to apply community detection methods to both the UK and Turkey cattle movement networks. Once again we must decide how to generate the network from the available data. Ideally we want to represent the network in such a way that captures the seasonal variation but does not aggregate on such a short time scale that the results are a consequence of noise rather than signal. To try to strike a balance we will aggregate the movements at a monthly level and at an annual level.

Not only will we consider aggregation of movements through time but we can also aggregate nodes together to reduce the size of the adjacency matrix and increase the run speed of the community detection algorithms.

As there are many parameters that can be varied, we thought it prudent to run the community detection algorithms on an aggregated network rather than on the full 70,283 farm network. The farms were aggregated to the 68 counties in the UK and a network was made from the inter-county movements, with intra-county movements appearing as self-loops.

4.2.1 County-level community detection

Firstly, the algorithm was run with the movements aggregated at an annual level. As we have only one year of data for the UK this is just a one-slice network so the coupling, ω, is irrelevant. We run the algorithm 1000 times and compute the VI between all pairs of partitions. The Q^{NG} for these partitions showed very little variation, with an average of 0.6278 with a standard deviation of 0.0013. The VI across all partitions ranged from 0 to 0.1868 with a standard deviation of 0.0525, thus we can conclude there is very little variation between different partitions (figure 4.4). Averaging across all the VI for every partition we find that 675 of the 1000 partitions all have the minimum average VI and inspecting these partitions we find they are all identical. This partition therefore becomes our definitive partition for the 2010 UK data aggregated at an annual level found by maximising Q^{NG}. We call this the annual County NG partition.

Plotting the counties on a map and colouring counties in the same community the same color we find that the UK is divided into six contiguous regions (figure

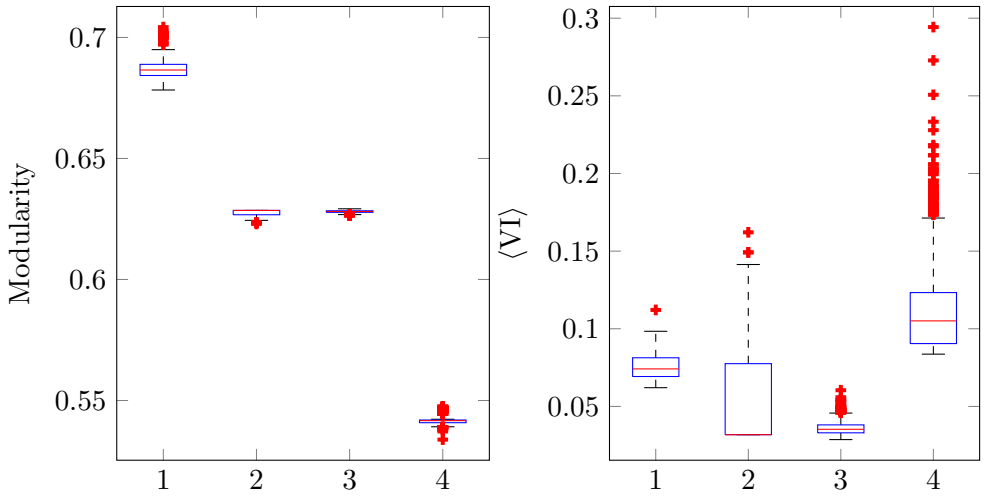

Figure 4.4: Box-plots depicting the modularity and average variation in information for different partitions and null-models for community detection in the UK. The partition with the lowest average VI for each was used as the seed partition for any further explorations with that method. 1 - Annual County NG partition, 2 - Monthly County with NG method and $\omega = 0$, 3 - Annual County SPA , 4 - Farm NG.

4.5) which are labelled as: Scotland, North England, North Wales, South Wales, East England, and South West England. The sizes of each of these communities in terms of number of counties and number of farms are given in table 4.1 as are the number of in, out, and internal movements.

The next step is to divide the annual data into monthly networks and search for communities on this time-dependent multi-slice network. The main parameter of interest here is the inter-slice coupling. At first we will set the coupling to zero; in effect this will create an independent community structure for each month. This is done 1000 times.

We now compare the full partitions to each other to find the most representative partition as we did for the annual partition. Upon calculating the average VI for each partition we see there is little variation between the partitions (figure 4.4). We choose the partition with the minimum average VI to be the representative partition. This partition has 6 communities for most of the year but increases to 7 communities for May, June and August.

Our next task is to add a non-zero coupling so we can track the evolution of the communities through time. As the coupling becomes encoded in the super-adjacency matrix its value is similar to that of a movement, there is no direct interpretation of this. With this in mind we experiment with different values of ω

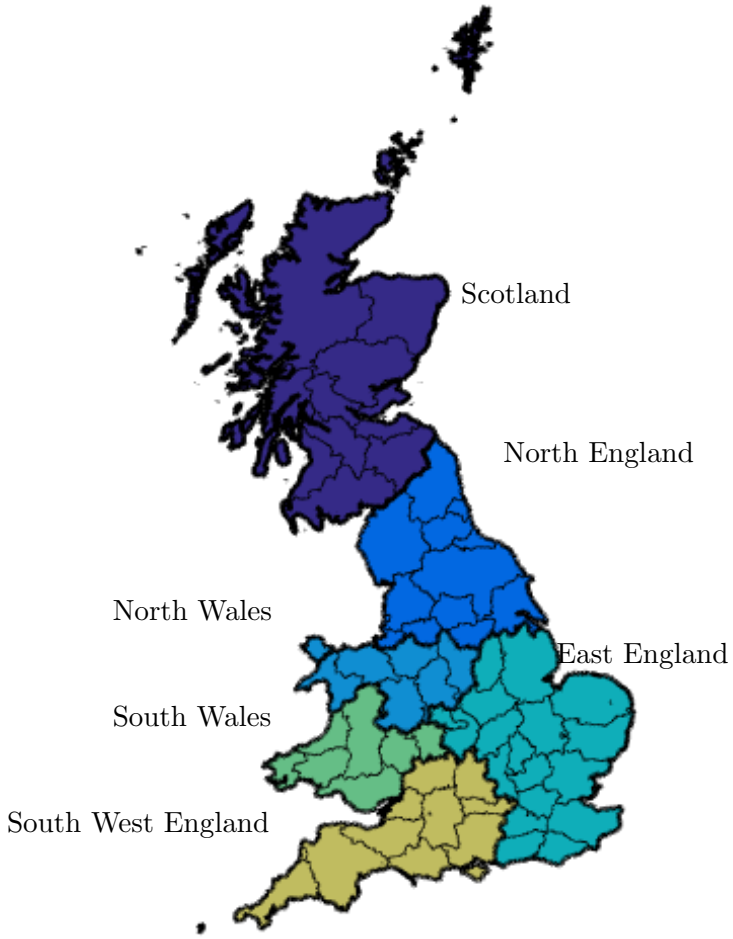

Figure 4.5: Communities found using the Louvain method to maximise Q^{NG} for the UK 2010 movement data with farms aggregated to a county-level. The 6 communities are given names that roughly match those of the regions where they are based.

Community	Counties	Farms	in	out	internal
Scotland	14	12016	82239	77318	63357
North England	12	13769	109236	129652	86336
North Wales	6	10771	95354	80748	57863
South Wales	7	9837	70733	54906	32660
East England	18	8570	38907	69495	25116
South West England	11	15320	101133	85483	73855

Table 4.1: Table summarising number of counties, farms, in-movements, out-movements and internal movements for each of the six communities found by running the Louvain community detection algorithm with the Newman-Girvan null model on the 2010 UK data aggregated at a county-level. These communities correspond to those shown in figure 4.5

.

to see if this has a strong effect on the community structure.

We run the community detection algorithm with $\omega = [0, 0.1, \ldots, 1, 2, \ldots, 3769]$ (3769 is the maximum weight of any of the edges in the 12 county-level adjacency matrices) using the seed found from the 1000 monthly partitions with $\omega = 0$. First we note that increasing the coupling increases the modularity (figure 4.6 A). While a high value of modularity is interpreted as evidence of a community structure, for the case of multi-slice community detection increasing coupling will always increase modularity if two partitions are equal as seen from equation 4.10. Therefore we should not assume the partition with the highest Q is necessarily the 'best' partition.

Calculating the average VI between partitions we see that for large coupling, $\omega > 350$ the partitions are all the same (figure 4.6 B). Indeed they are the same as the annual partition, and examining their sub-partitions (divisions into months) we see that they are temporally static (figure 4.6 C). This is not surprising as the coupling term is clearly dominating the modularity. On the other end of the scale for small coupling ($0.1 < \omega < 2$) all partitions are the same but they do have temporal variation.

The structure found for the $\omega = 0.1$ partition is, on a monthly basis, the same as that for $\omega = 0$ partition in effect tracking the evolution of the community structure through time. For the remaining values of ω in the range considered the mean VI starts high and begins to decrease as we get closer to the static partition found at high coupling, but of course this is not surprising as over 3000 of these partitions are the same and hence will contribute zero to their mean VI scores. Indeed, of the 3789 partitions created only 97 of them are unique. The variation

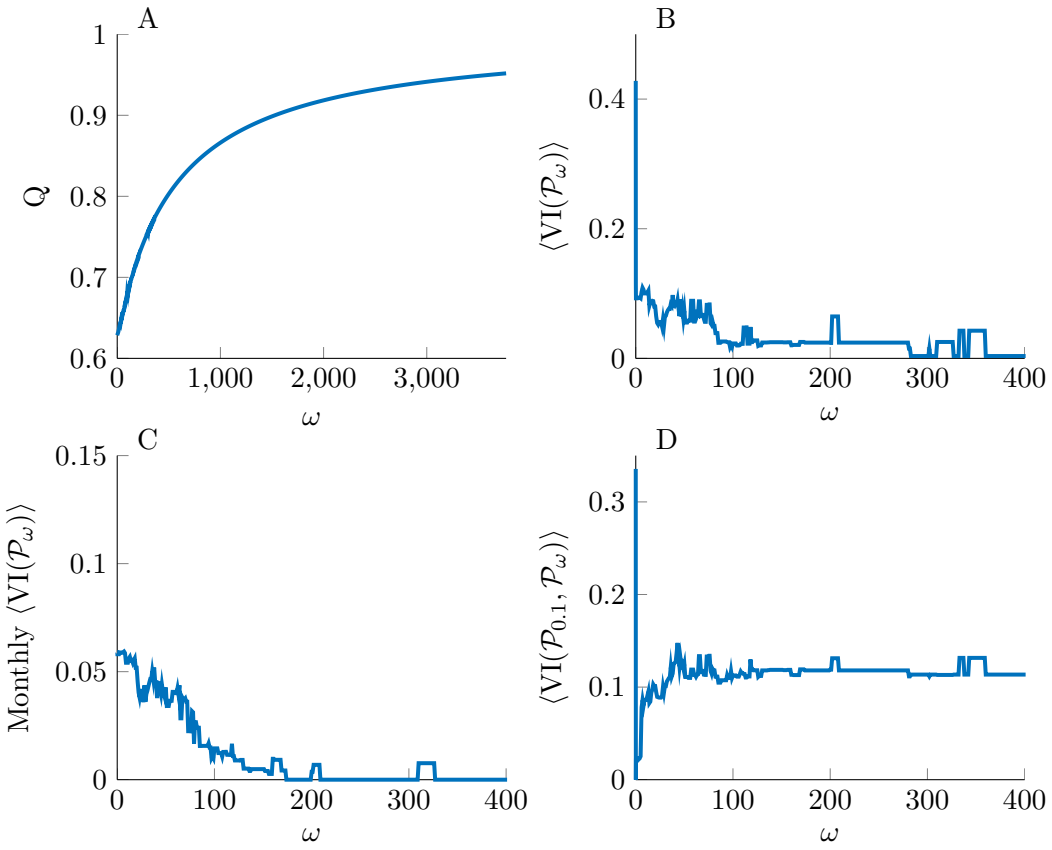

Figure 4.6: Plots of the properties of the community structures of the county-level cattle network for the UK with monthly aggregation through 2010 as the coupling ω is increased. A) is the modularity, B) the mean variation in information between all partitions, C) the mean variation in information between the sub-partitions (monthly partitions) for each partition, D) the variation in information between every partition and the $\omega = 0.1$ partition.

between these partitions and the $\omega = 0.1$ partition is also examined in figure 4.6 D and is fairly small.

The VI between all 97 unique partitions is small therefore one would not expect an individual analysis of these partitions to yield particularly insightful results. Therefore we will will limit ourselves to considering the partitions found for $\omega = 0.1$ and the static partition found for high coupling values[ii].

The evolution of the community structure through time is shown in figure 4.7. The annual community structure (with the same colour scheme as figure 4.5) is shown on the left with the nodes reordered to keep the communities in distinct groups. The community structure for each month with $\omega = 0.1$ is then shown

[ii]The static partition is the same as the annual partition

evolving through time. The communities vary slightly from those found in the annual network particularly in May, June and August and November when counties originally designated in the East England group merge into the community initially designated as being North Wales (figure 4.8). The North of England stays fairly stable throughout as does Scotland, except in August (figure 4.8 D) when the county of Na h-Eileann an Iar becomes isolated from the rest of Scotland but still connected to the previous instances of that community.

As modularity is additive (equation 4.12) we can directly calculate the modularity from each community and the change in modularity from moving a county between communities. Calculating the modularity for the whole of Scotland throughout the year and the contributions from the two separate communities, we find that separating Scotland into two communities does marginally increase the overall modularity. We also calculated the modularity if Na h-Eileann an Iar were to be a separate community for the month of August with no link to the previous community in Scotland. Interestingly this would actually have increased the overall modularity more than the given partition but this gain was so small as not to pass the threshold required by the Louvain method to be considered a worthwhile division.

The fracturing of Scotland into two communities comes from Na h-Eileann an Iar having very few movements, only 1,481, throughout the year. Almost a quarter of these movements, 365 were within-county, the majority were out-movements, 995 with the remaining 121 being in-movements. In August however, Na h-Eileann an Iar was only involved in 31 movements, 29 of which where internal. Thus Na h-Eileann an Iar was effectively isolated from the rest of Scotland in August, resulting in the previously described modularity values.

We analyse the evolution of the community structure further by defining the mode partition for the UK with $\omega = 0.1$. Each county is assigned to the community it appears in most often throughout the year (figure 4.9). This partition is quite similar to the annual partition with VI between the two being 0.057. The communities of Scotland and North-England are the same for both partitions. The South West of England and the East of England also change little with only the county of Oxfordshire moving between the communities in the two partitions. Wales and some of its bordering counties that go through the most variation as seen from the maps of figure 4.8.

4.2.2 Farm-level community detection

The county-level community detection has given us some promising results. However, we would like to probe deeper into the community structure of the network

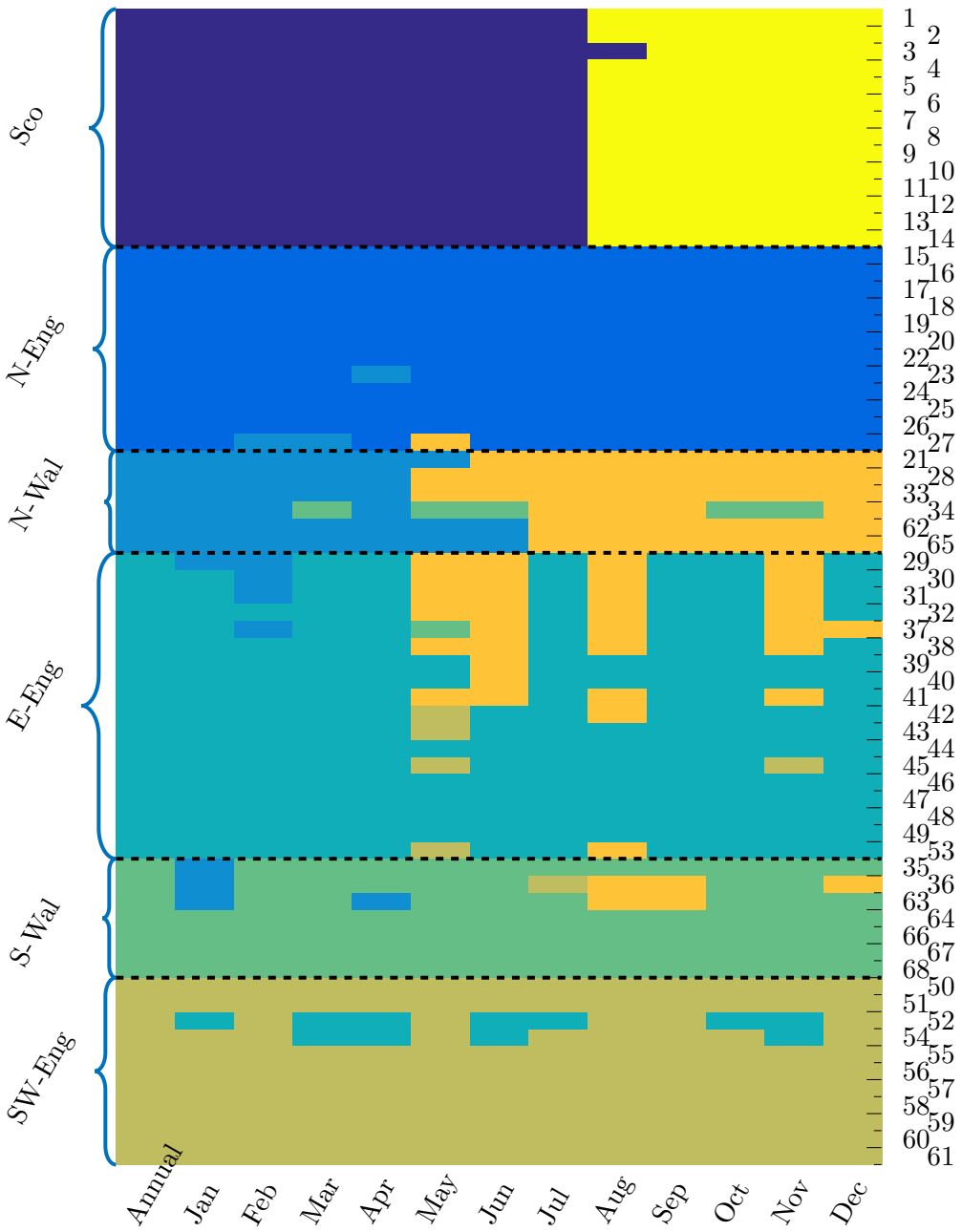

Figure 4.7: Evolution of the community structure of the county-level cattle network for the UK with monthly aggregation through 2010 found using the Newman-Girvan null-model with a temporal coupling of $\omega = 0.1$. Colours indicate counties that are in the same community with the initial grouping corresponding to the community structure found at the annual aggregation. The counties are reordered so that the colour blocks for the annual structure are contiguous. The individual counties are labelled in accordance with table C.1.

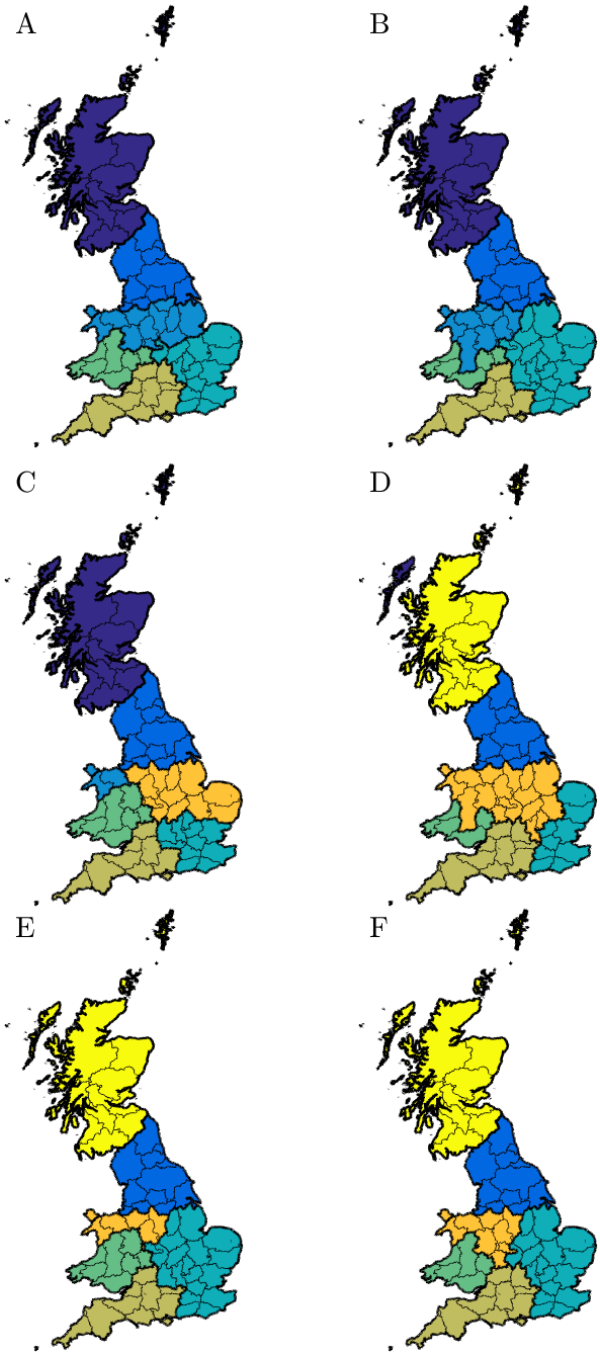

Figure 4.8: Maps showing the community structure of figure 4.7 for the months of A February, B April, C June, D August, E October, and F December.

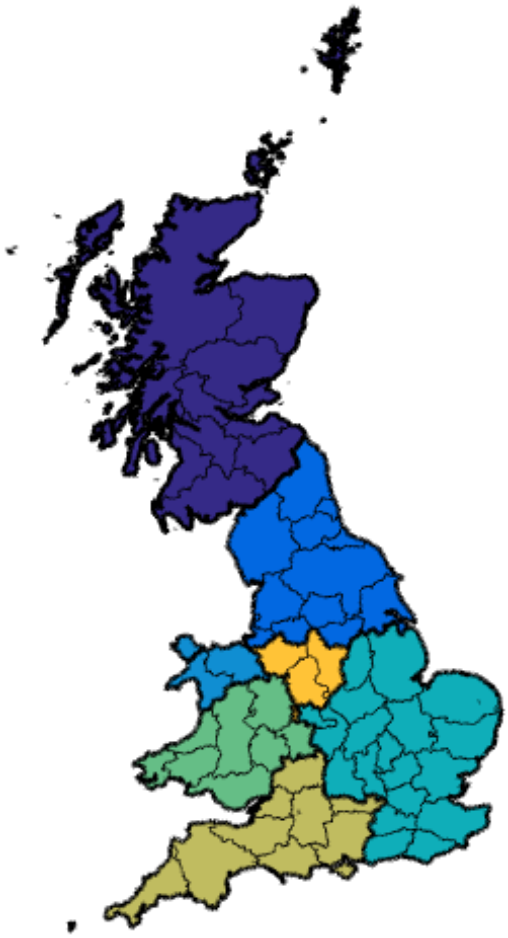

Figure 4.9: Map showing the mode partition found from the evolution of the UK community structure with a coupling of 0.1.

of individual farms to try to gain a deeper insight into the farm-level interactions of the cattle network. We can also compare our results with community detection performed by Kao et al. [2006] on the network of sheep movements in the UK. They found highly regionalised communities with Scotland and the North of England being particularly distinct.

The algorithm detects 352 communities. Most of these are small - 331 have a size of less than 10 farms. The two largest communities have 12,332 and 11,667 farms which is 35% of the total number. Figure 4.10 shows the location and community of each farm in the ten largest communities in the partition. This plot is difficult to interpret due to the high density of farms in the centre of the country which causes some of the points to be obscured.

In order to better understand the spatial impact of the community structure we drew polygons around the densest clusters of farms from each of the 10 biggest communities. Each polygon was coloured different for each community. If polygons from different communities overlapped this was indicated by cross-hatching the overlapping area (figure 4.11). The precise procedure of how this figure was made is given in the appendix C.

Some aspects of the community detection at the county level are reflected in the farm level map, Scotland for instance has very little overlap with England, though it is now divided into two communities, Aberdeenshire and the islands, and the south of Scotland. The farms in the east of England also form a distinct community. The most striking difference is the large overlapping community of 7,384 farms that is centred at the meeting points of Wales, The Midlands and South West England. Another feature that was not present in the county data is the non-contiguous community of South West Wales and South West England.

4.2.3 The spatial null model

While extremely useful and applicable to many different types of networks the Newman-Girvan null-model relies purely on network topology and has no facility to incorporate a network embedded in space. If the main driver behind connections being formed is spatial proximity then the edges in the communities found will have a highly correlated spatial dependence. By incorporating space into a null-model we would hope to find communities that are space-independent. As farm location data is available to us we wish to use a null-model in our analysis that can find such space-independent communities. Such a model should reward edges in communities that appear at distances we may not expect to see them and penalise edges that appear at frequently observed distances. For many spatial networks long distance

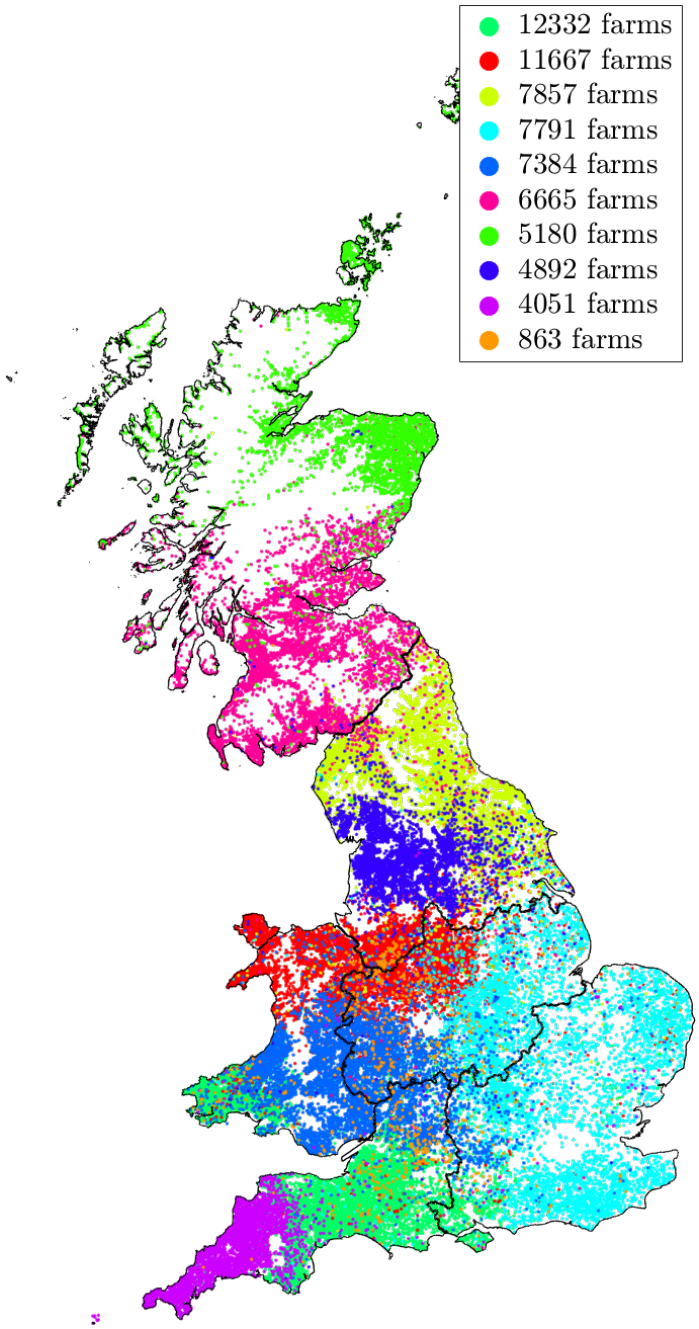

Figure 4.10: Map showing the farms of 10 the largest communities found in UK cattle network. Overlaps in area are cross-hatched and the borders of the six previously defined region of the UK are drawn in bold.

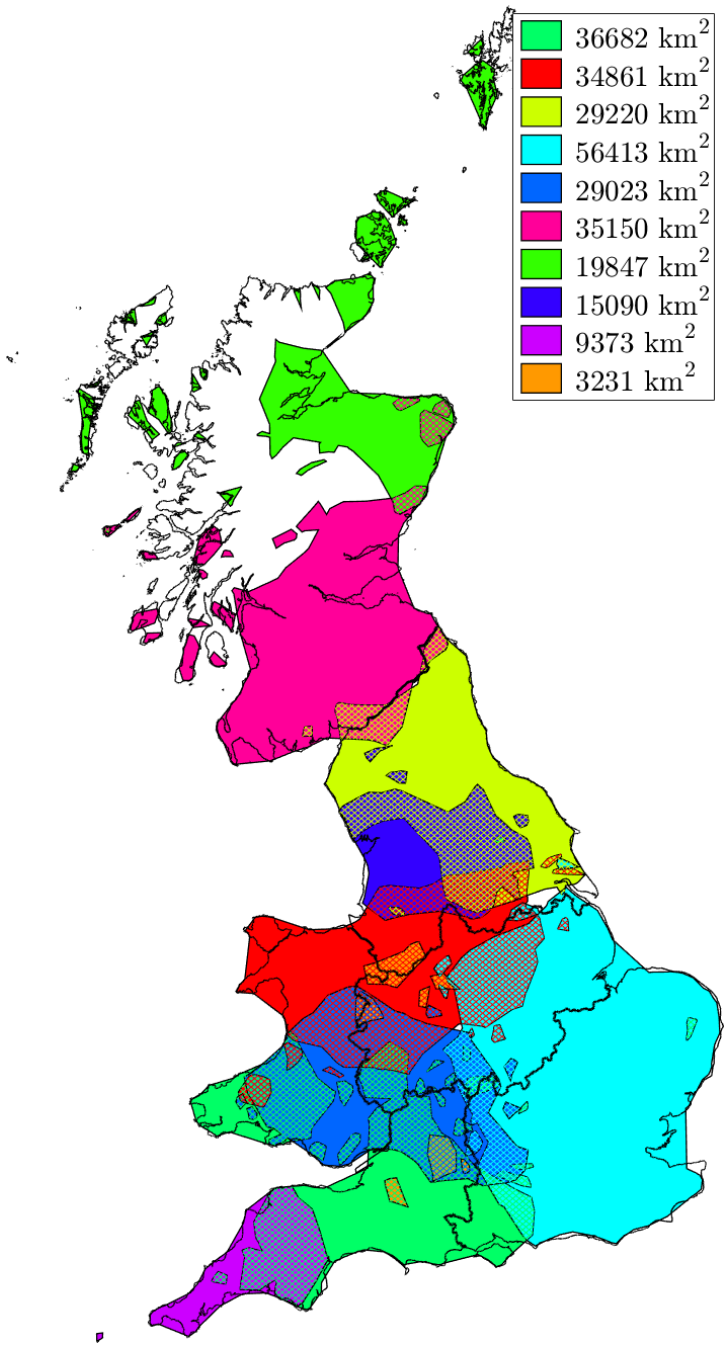

Figure 4.11: Map showing the area covered by the largest communities found in UK cattle network. Overlaps in area are cross-hatched and the borders of the six previously defined region of the UK are drawn in bold. The area of each community is given in the legend.

edges are rare as they are typically costly, thus if they do appear this may yield some information about the two nodes connected by the long distance edge that the adjacency matrix alone does not reveal[Barthélemy, 2011].

Telecommunications networks between cities [Krings et al., 2009], human migration [Levy, 2010], and the International Trade Market [Bhattacharya et al., 2007] have all been successfully modelled by the interaction matrix $\mathbf{T}$

$$T_{ij} = N_i N_j f(d_{ij}) \tag{4.26}$$

where T_{ij} represents the volume of interaction between node i and j, N_i is a measure of importance of node i (for example its population), and f describes the influence of space on the interaction as governed by the distance d_{ij} between the nodes.

The spatial function f is typically fitted from data and it is generally found that f is inversely proportional to the distance raised to some power $f(d_{ij}) \propto d_{ij}^{-\alpha}$. These models are typically referred to as gravity models as the roles of distance and importance play analogous roles as distance and mass in Newton's law of gravity.

The model proposed by Expert et al. [2011] builds on the idea of a gravity model but does not presume a form for the role played by distance in f but instead allows it to be determined entirely by data. Their null-model is

$$P_{ij}^{\text{Spa}} = N_i N_j f(d_{ij}) \tag{4.27}$$

where N_i is equivalent to that of N_i in equation 4.26 and the form of f is

$$f(d) = \frac{\sum\limits_{i,j|d_{ij}=d} A_{ij}}{\sum\limits_{i,j|d_{ij}=d} N_i N_j} \tag{4.28}$$

and is the weighted average of the probability $A_{ij}/(N_i N_j)$ for a link to exist at distance d. Summing over P_{ij}^{Spa} will give the total weight of the network.

The importance of each node can be set as the population of an area or the number of farms in a district; the node's degree could also be used. If this is the case and all edges are equally likely at any distance, i.e $f(d)$ is constant then the Newan-Girvan model is recovered. We now refer to the modularity function that uses this spatial null-model as Q^{Spa}.

It is important to note that while $f(d)$ is driven entirely by the data supplied, some binning of the distances involved will have to be taken to give meaningful values for the probabilities and to keep the size of the function manageable. The distributions of distances at various bin widths will inform the choice of bin width

for $f(d)$.

When searching for partitions we have considered the Newman-Girvan modularity and the spatial modularity separately. However, Expert *et al* have suggested a method to interpolate between the two null-models by using a mixing parameter ξ such that

$$P_{ij}(\xi) = [\xi P_{ij}^{\text{Spa}} + (1-\xi)P_{ij}^{NG}]. \tag{4.29}$$

We extend the modularity to the spatial null-model for a directed, weighted network as

$$Q^{\text{Spa}} = \frac{1}{w}\sum_{i,j}\left[A_{ij} - N_i^{\text{in}}N_j^{\text{out}}\frac{\sum\limits_{i,j|d_{ij}=d} A_{ij}}{\sum\limits_{i,j|d_{ij}=d} N_i^{\text{in}}N_j^{\text{out}}}\right]\delta_{c_i,c_j} \tag{4.30}$$

if population size, or any equal measure, is being used as the measure of importance then $N_i^{\text{in}} = N_i^{\text{out}} = N_i$ and we retrieve the original definition of Q^{Spa} [Expert et al., 2011].

4.2.4 UK community detection with the spatial null model

We now repeat the work of the previous section but use the spatial null-model given by equation 4.30. We again aggregate to the county-level.

For the importance measure in the null-model we will use the number of farms in each county. The spatial-null model requires a discrete set of distances as it deals with the probability of an edge to occur at a certain distance. We take the distances between counties to be the distances between their centroids and experiment with different bin widths on the annual data.

We test bin widths at distances of 1 to 10 km and at 15,20,25,50,75,100 km. For each bin width we run 1000 different instance of the community detection algorithm and compare the average VI across all partitions to find the most representative partition for that bin width.

The average VI between all the representative partitions is then found, the minimum of which gives us the representative partition for the most representative bin width, this occurs at 6km (figure 4.12 A). We also plot the deterrence function (equation 4.28, figure 4.13) at the various bin sizes. The effect of increasing bin width is to smooth out the function at large distances but at the expense of losing detail at distances near the bin width.

The structure found using the spatial null-model varies greatly from that found using the NG null-model. Looking at the annual structure first; for the

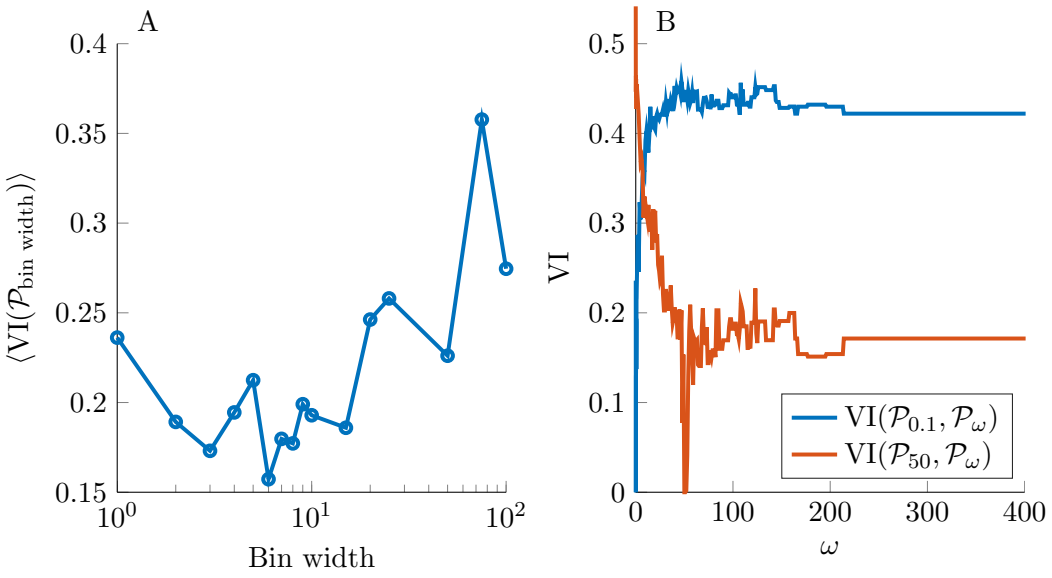

Figure 4.12: Plots depicting the key steps in finding the most representative bin width and inter-slice coupling ω using the spatial null-model at the county level. A) The average VI between the representative partition found across a range of bin widths. B) The VI for the $\omega = 0.1$ and $\omega = 50$ partitions between partitions found with a range of coupling. Each partition was found using a bin width of 6km.

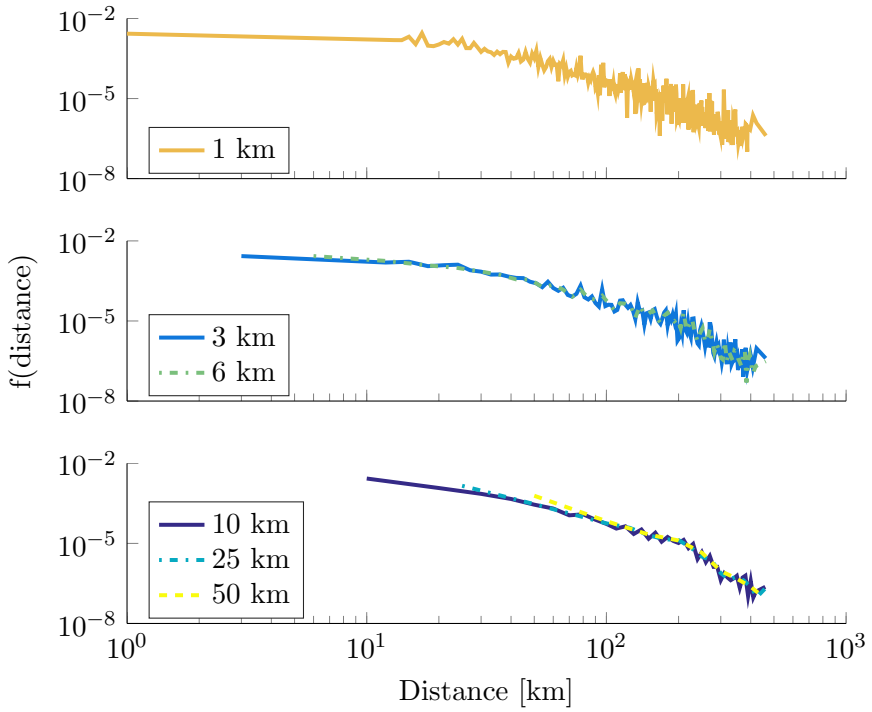

Figure 4.13: Plots of the deterrence f function as distance is varied calculated from the 2010 UK data aggregated at a county level for different bin widths.

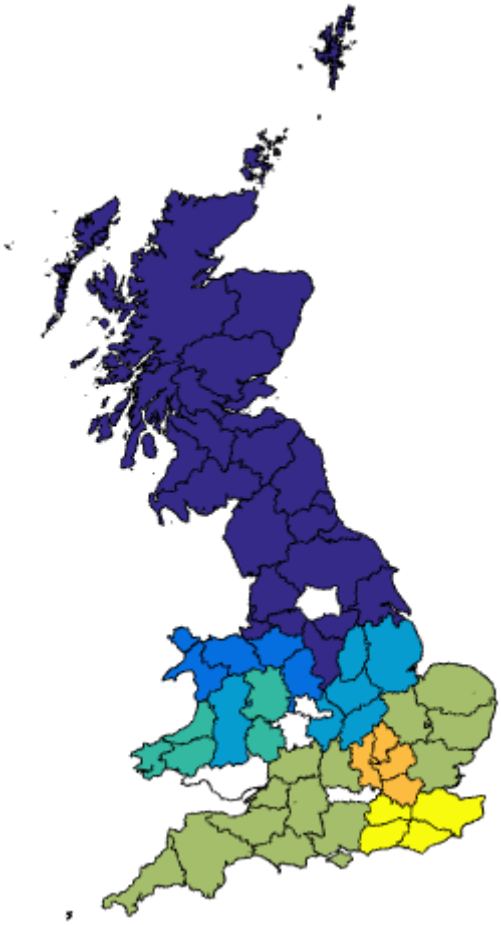

Figure 4.14: County-level map depicting the communities found using the spatial null-model with a bin size of 6km with movements aggregated to an annual level, communities of size one are coloured white.

NG model we found 6 communities and though the number of counties in each community did vary the number of farms in those communities was around 10,000 for each community. However, for the spatial null-model there were 12 communities (figure 4.14) and the distribution of the number of farms in each community was far more varied, there were two dominant communities in excess of 15,000 farms, three communities with 5,000 farms with the remaining communities ranging from sizes of 100 to 2,000 farms, table 4.2.

These results are in line with those found by Expert *et al* for the Belgian mobile phone network for which they implemented the spatial null model, yielding a heterogeneous community structure dominated by two large cliques compared to more homogeneous structure found using the NG model.

Community	Counties	Farms
Scotland & North England	26	26430
North Wales	4	7115
West Wales & West Midlands	5	7236
East Midlands & Powys	6	5993
South East England	4	1949
South West & East England	14	17079

Table 4.2: Table summarising the six largest communities found by running the Louvain community detection algorithm with the spatial null-model on the 2010 UK data aggregated at a county-level with a bin size of 6km.

As we introduce inter-slice coupling for the monthly networks the results are not as clear cut as for the NG case. Where before we found that the $\omega = 0.1$ partition was able to track the evolution of the zero coupling partition through time, no equivalent value of ω could be found for the spatial model. The average VI between all unique partitions reveals the $\omega = 50$ partition to be the most representative. The VI($\mathcal{P}_{0.1}, \mathcal{P}_\omega$) and the VI($\mathcal{P}_{50}, \mathcal{P}_\omega$) are plotted in figure 4.12 B. We look at the evolution of $\mathcal{P}_{0.1}$ and $\mathcal{P}_{50}$ in detail in figure 4.16. Here we see that for $\mathcal{P}_{0.1}$ the structure is not very coherent with new communities being formed monthly even for the large group of counties in Scotland and North England. There are also many communities of only one county and counties change community many times (figure 4.15). For the $\omega = 50$ partition the higher coupling links the communities more strongly through time and we see more stable communities (figure 4.16) with counties being assigned to at most two communities (figure 4.15). We also see that the spatial null-model places East Anglia in the same community as the South-West. Both of these regions have a high cattle density, which the NG null model did not pick up.

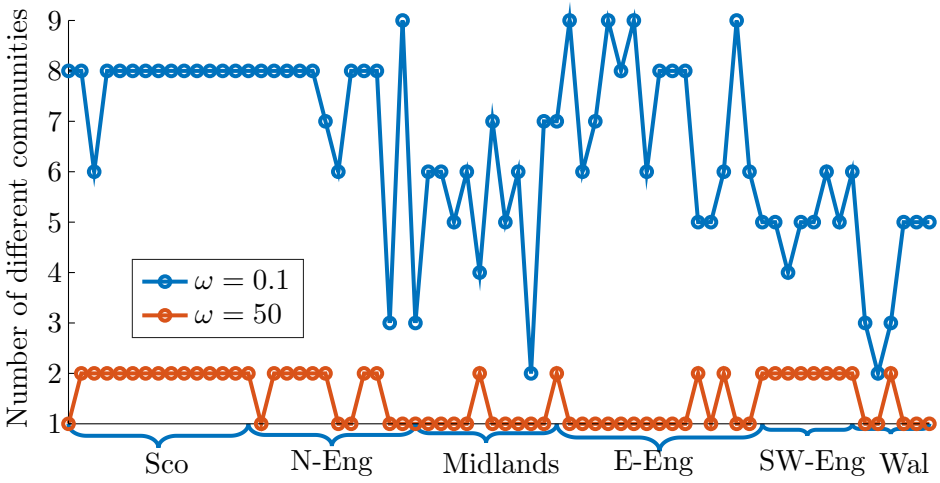

Figure 4.15: The number of different communities that each county is assigned to using the spatial null model for the $\omega = 0.1$ and the $\omega = 50$ partitions.

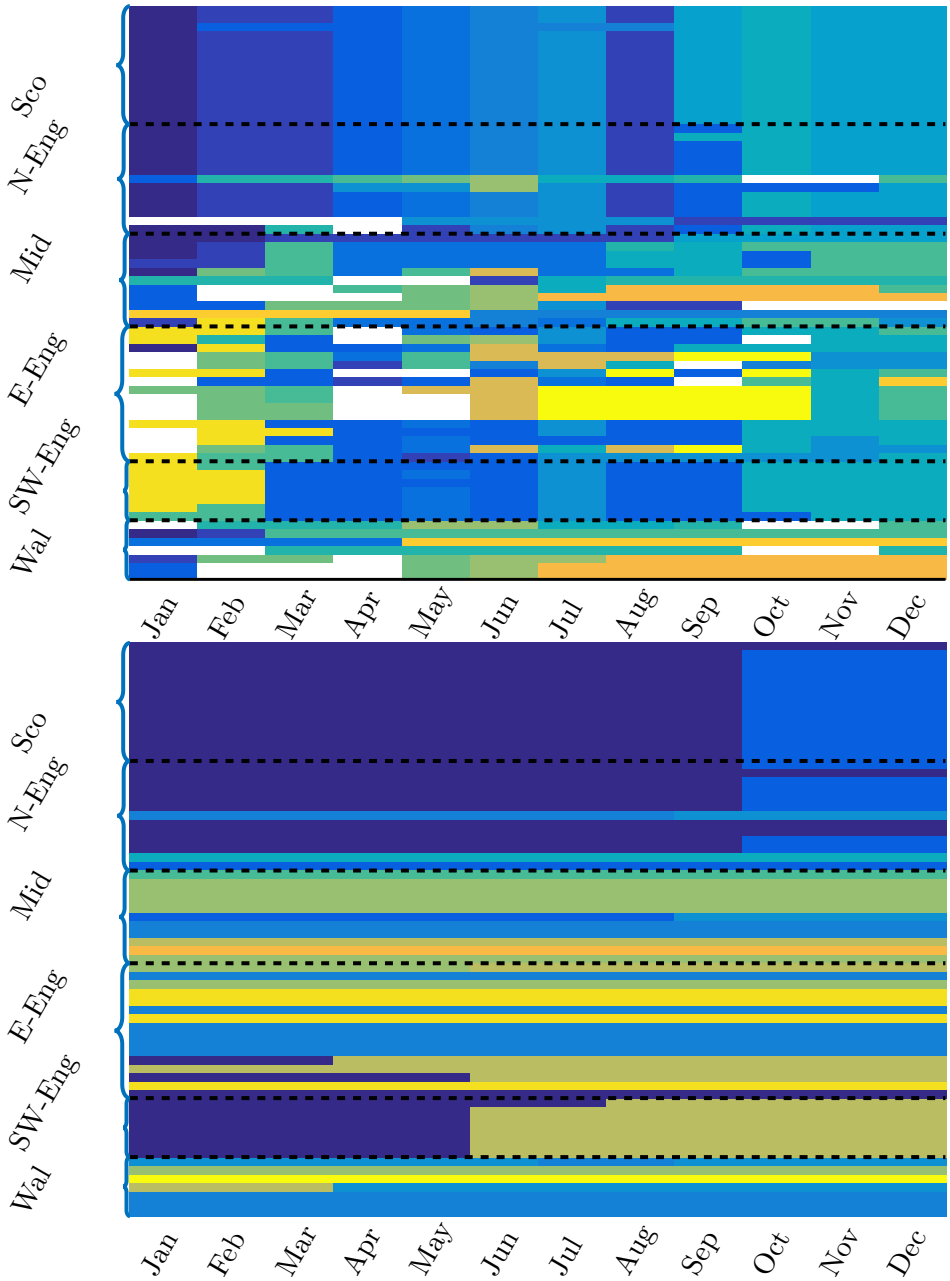

Figure 4.16: Monthly county level community structure found using the spatial null model with a bin-width of 6 km, the inter-slice slice coupling is $\omega = 0.1$ (upper) and $\omega = 50$ (lower). The counties are grouped into the geographical regions of Scotland(Sco), North England (N-Eng), the Midlands (Mid), East England (E-Eng), South-west England (SW-Eng) and Wales (Wal). Communities having less than one county per month are coloured white.

4.3 Community detection in Turkey

Turkey can be resolved at province, district or farm level. There are 81 provinces, 928 districts and 55,000 epi-units. Turkey has an area of over 780,000 km^2 whilst the UK excluding Northern Ireland is only 230,000 km^2 in area. This makes the administrative divisions amongst countries hard to directly compare as each province in Turkey has an average area of 9,600 km^2 and each district an area of 844 km^2 while each county in the UK has an average area of 3,400 km^2.

We begin with the Turkish data from 2010 and analyse the community structure at a provincial level using the Newman-Girvan null model. We find there are 17 communities which are made of contiguous provinces except interestingly for one community, Ankara in central Anatolia is placed in the same community as Ezurum and four of its neighbouring counties (figure 4.17). Ezurum has the highest number of cattle movements of any province. Both of these provinces were identified as areas of disease persistence in the analysis of Gilbert et al. [2005].

Figure 4.17: Provincial level community detection in Turkey. Communities are coloured distinctly and made of contiguous blocks except for the one community marked with a star which is made of two separate blocks of provinces.

At district level the map of Turkey looks very similar to the provincial one (figure 4.18) and the variation in information between the two partitions is relatively low at 0.11. Some districts move from their provincial community but the community blocks remain largely contiguous save for the strong Ankara - Ezurum connection which now includes the district of Yeşhilisar in the Kayseri province. The districts of Istanbul are now split by the Bosphorus with the European districts joining the community containing the provinces of Thrace. The city districts of Istanbul are assigned into 5 different communities. This splitting is expected as the Thrace region has far stricter movement controls, while the city districts will have imported cattle form across the country.

The temporal evolution of the community structure of Turkey appears less

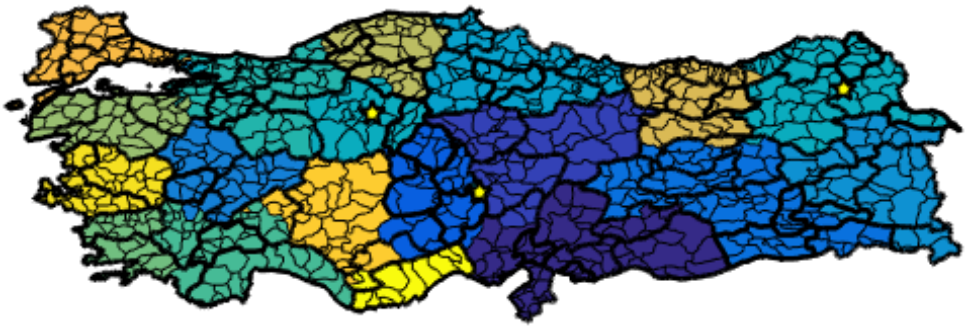

Figure 4.18: District level community detection in Turkey. Communities are coloured distinctly and the provincial boundaries are drawn in bold to allow for comparison with figure 4.17. The starred regions all form the same community.

stable than that of the UK; the average VI between months in the same partition is higher for Turkish partitions than for UK ones with the same coupling (figure 4.19 A). The number of communities found at the annual level in Turkey is three times that of the UK which allows for more variation once a monthly coupling is introduced.

This variation and the large number of communities in Turkey, 35 and 20 for $\omega = 0.1$ and $\omega = 50$ respectively compared to 7 and 6 for the same coupling values in the UK (figure 4.19 B), make the visual interpretation of the community structures evolution difficult. As an example we plot the evolution of Turkey in two ways using the high coupling [iii]. The evolution of the community structure is plotted where the provinces are grouped into the 7 geographical regions of Turkey (figure 4.20) and grouped by the communities found in January 2010 (figure 4.21).

4.3.1 Farm level community detection in Turkey

For the full farm level we have coupled the years from 2007-2012 through time with a weak coupling of $\omega = 0.1$. Maps for each year are seen in figure 4.22 and figure 4.23. This was only carried using only one coupling value as due to the size of super-adjacency $(55,193 \times 6)$ run time was long (≈ 3.5 days).

We identified 21 large communities while the largest small community had a size of 6 epi-units, which would correspond to one farm being in its own community for all six years. In total the 21 large communities accounted for over 90% of the farms, and ranged in size from 32,000 farms to 3,000 farms. The sizes of the communities are fairly consistent save for some of the communities such as 18-21 which only become large in later years, figure 4.24.

[iii] Low coupling gives rises to too many communities to be able to distinguish the colours.

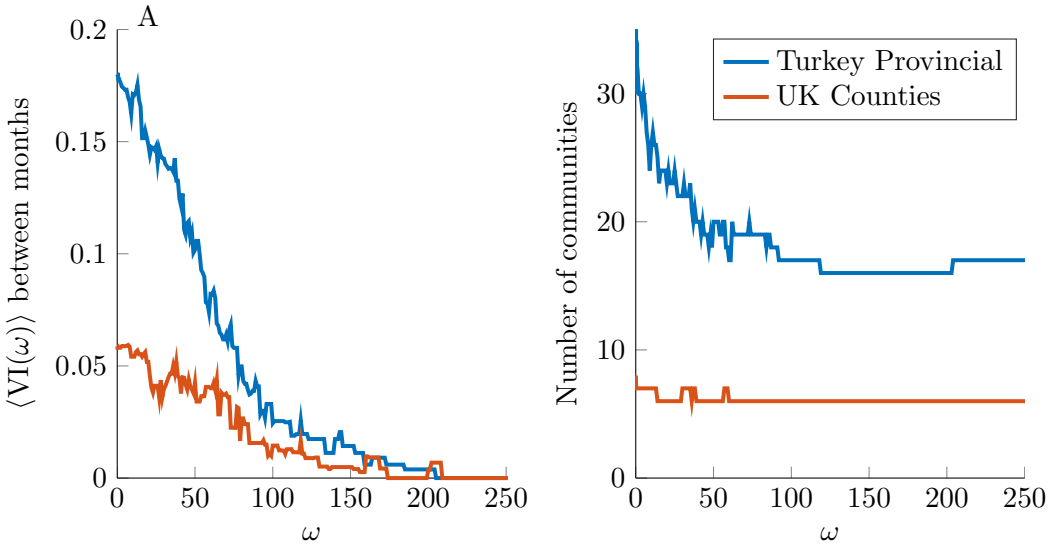

Figure 4.19: Plot A shows the average VI between the months for each partition found as coupling is increased. Plot B shows the number of communities identified in each partition as coupling is increased. Both plots show results for Turkey at a provincial level and the UK at a county level.

The link between Ankara and Ezurum (community 4) which we had identified in the district and provincial community structures for the 2010 data is also present at a farm level but only for the first four years examined, by 2011 this strong connection has disappeared. The size of community 4 centred around Ezurum undergoes a large drop in size in 2011 suggesting that the connection between the Ezurum province on the surrounding provinces in North East Anatolia has weakened.

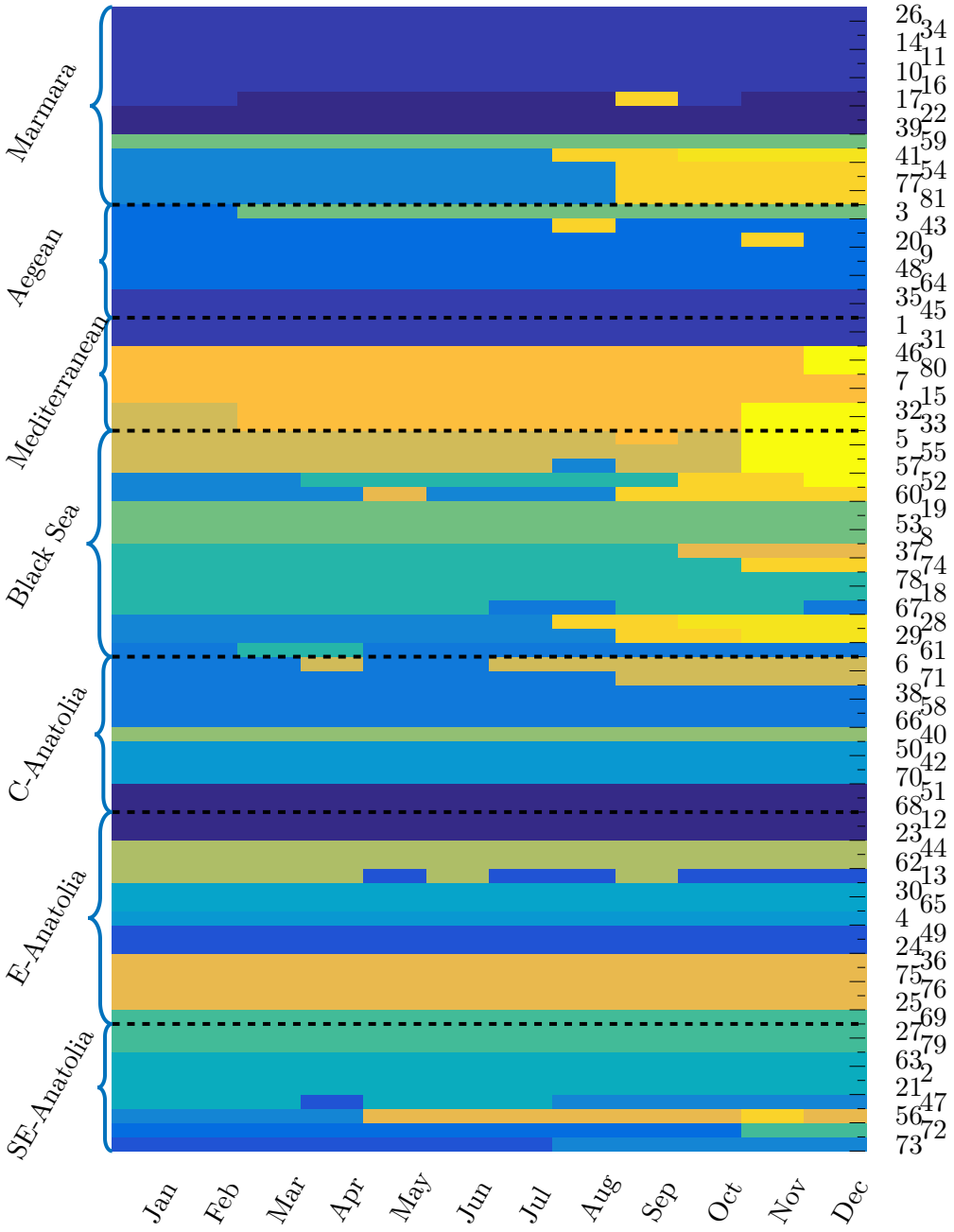

Figure 4.20: Evolution of the provincial level community structure of Turkey for 2010 with $\omega = 50$. The provinces are grouped by regions and the labels correspond to table C.2 in appendix C. There are 20 different communities.

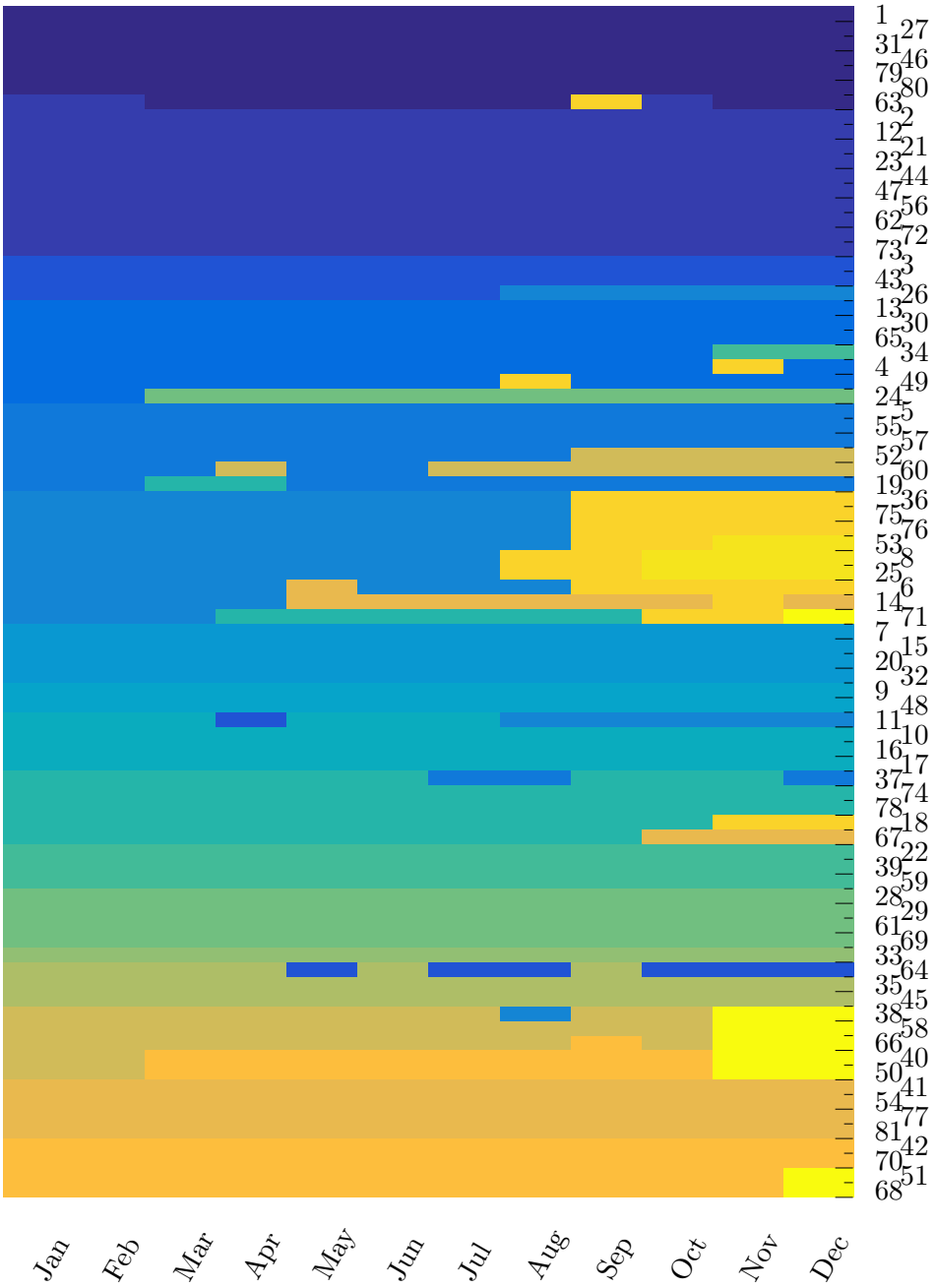

Figure 4.21: Evolution of the provincial level community structure of Turkey for 2010 with $\omega = 50$. The provinces are grouped by community and the labels correspond to table C.2 in appendix C. There are 20 different communities.

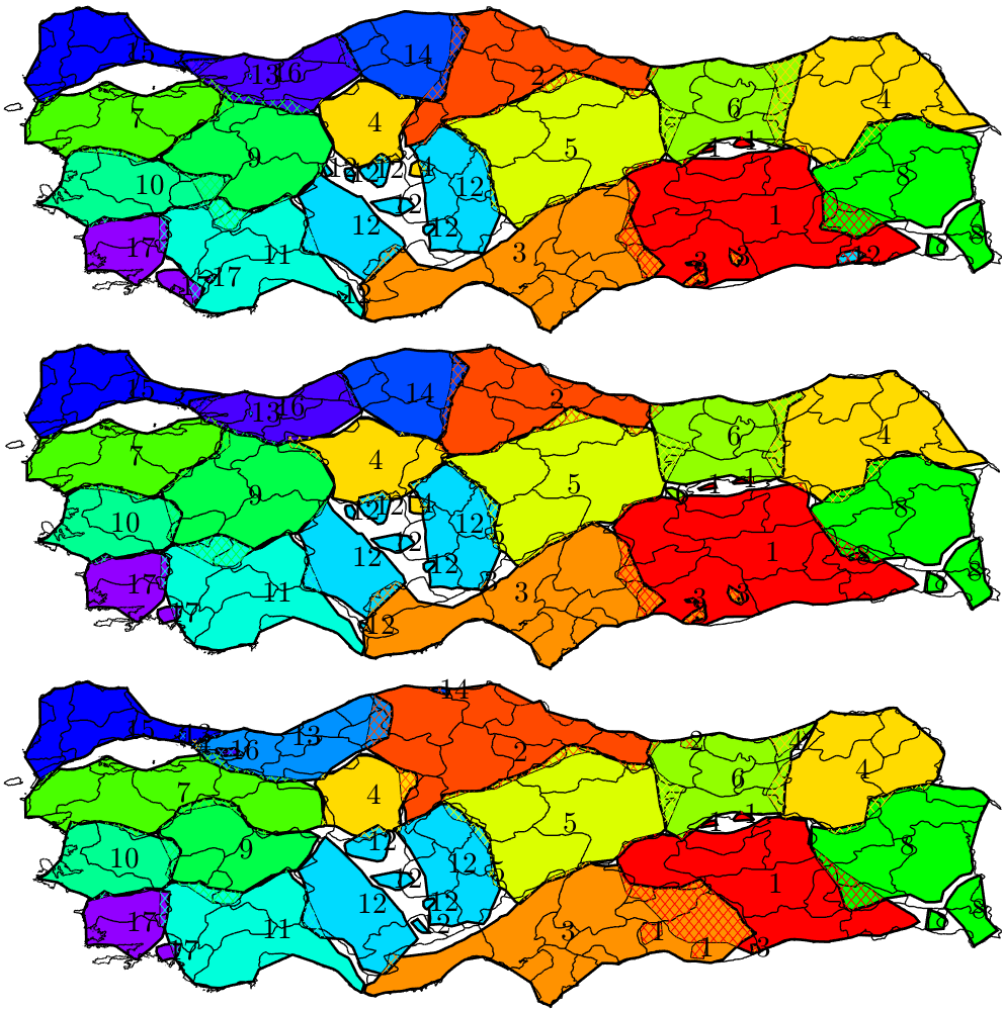

Figure 4.22: Coupled farm level community detection for Turkey. From top to bottom: 2007, 2008, 2009. Maps show the 21 biggest communities with cross-hatching indicating areas of community overlap.

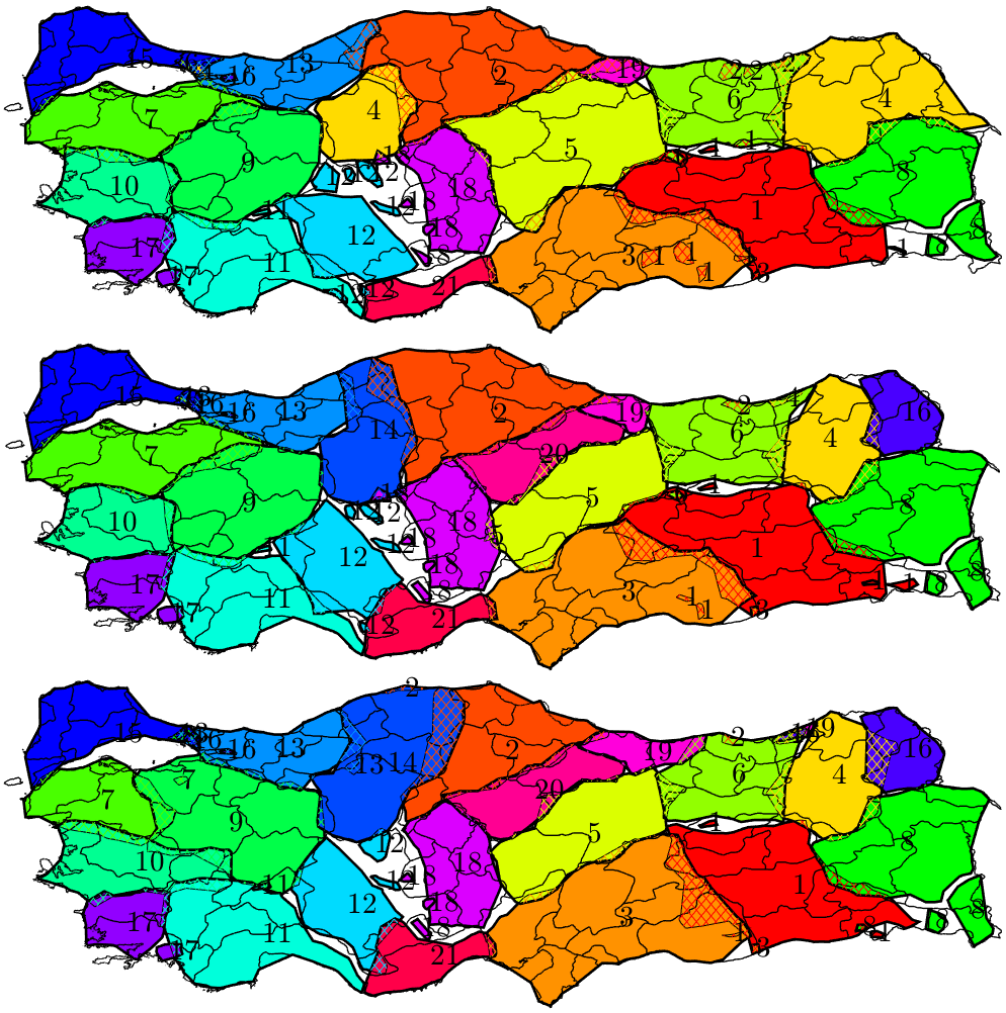

Figure 4.23: Coupled farm level community detection for Turkey. From top to bottom: 2010, 2011, 2012. Maps show the 21 biggest communities with crosshatching indicating areas of community overlap.

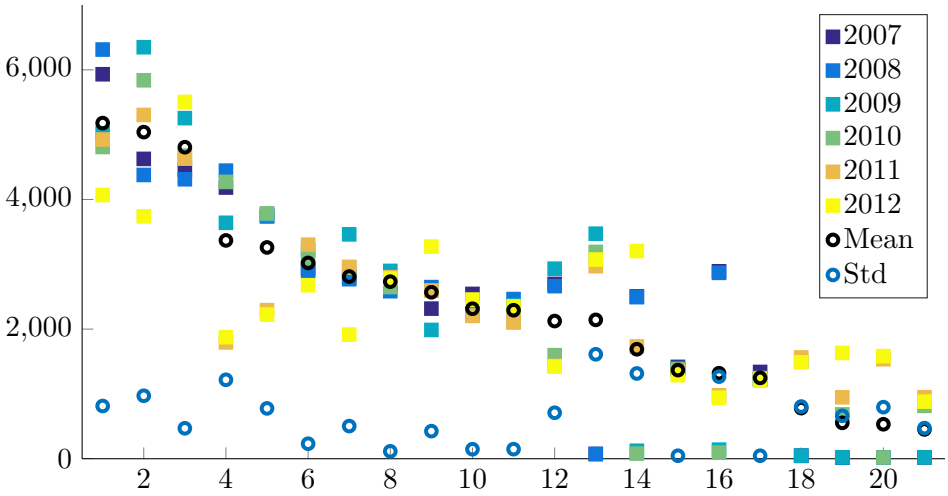

Figure 4.24: The sizes for each of the 21 large communities identified from 2007 to 2012 for the Turkey community detection across the six years with a coupling of $\omega = 0.1$. The mean size and standard deviation are also plotted.

4.4 Implications to epidemic prediction

Work by Salathé and Jones [2010] showed that the value of modularity found for a network can influence epidemic dynamics across the network. This work was carried out in a controlled manner by generating 50 small-world style sub-graphs and then connecting them together with inter-community links. These links were then rewired so that the modularity could be tuned, allowing for a variety of community structures. Community structure acted in two ways: strong community structure either served to stifle the spread of the disease with the epidemic becoming trapped in one community, or led to longer epidemics as the disease spread through communities in serial.

The focus of our work is slightly different from Salathé and Jones [2010]. In order to completely describe the effect of community structure in the cattle network of the UK we would have to be able to create networks with the same degree distribution as the original network but with some mechanism that would allow us to vary the modularity. To our knowledge there is no procedure that allows for the manipulation of the community structure of a network with an arbitrary degree distribution.

Instead we will focus on community detection as a predictive tool for the spread of epidemics across the UK. We will once again focus on the farming hot-spots of Cumbria, Devon, and Aberdeen and seed epidemics in each of these counties. The county-level community detection algorithms revealed that these counties appear in communities of varying levels of temporal stability (figure 4.7). While at the farm-level farms from these counties appear in communities that have varying degrees of overlap between other communities (figure 4.11).

As we learned from chapter 3 outbreaks seeded in different parts of the country can have different patterns of spatial spread. Epidemics in Cumbria tended to be localised to the North of England with some spread extending into Scotland and into the Midlands. Epidemics in Aberdeen were strongly regionalised staying almost entirely in Scotland. Outbreaks seeded in Devon tended to spread all over the country. Figure 4.11 somewhat reflects this - there are two communities in Scotland, one centred around Aberdeen and the other comprising the south of Scotland that have very little overlap with any other communities. This, coupled with the temporal and static county level partitions found which always grouped the Scottish counties together, show that the cattle network in Scotland is very close knit.

From figure 4.11 we see two large communities in the North of England, but unlike the Scottish communities they share a large overlapping region and also over-

lap with the communities in the Midlands. Devon is part of the largest community identified with the farm-level NG comprising 12,332 farms and overlapping with other large communities in Wales and central England.

With these observations about the community structure and our previous knowledge of disease dynamics within the cattle farms of the UK, we would expect to be able to predict the spread of disease for epidemics seeded in Aberdeen to a high level of accuracy. However, the expected results for Cumbria and in particular Devon are less clear cut.

For this investigation we consider farms to be our basic epidemiological unit and ignore within herd dynamics. This is the most widely used approach for national level livestock models [Kao et al., 2006; Keeling and Eames, 2005; Keeling et al., 2001], simplifying the model and increasing simulation speed. This is also the approach used in chapter 3. As we are only interested in epidemics that take-off we once again set a take-off threshold of ten infected farms and continue simulations until we have 5000 outbreaks that have reached this threshold in each of the three counties.

The probability of infection is again given by

$$P(S_i \to I_i) = 1 - \exp\left(-\beta \sum_{j \in I} A_{ij}\right) \tag{4.31}$$

and each farm stays infected for an infectious period of T days after which it is recovered. We vary $\beta \in \{1, 2, 3, 4, 5\}$ day^{-1} and $T \in \{7, 14, 21, 28\}$ days. Each simulation runs until the epidemic dies-out. We record the size of the epidemic and the farms infected at each daily time-step. We then compare the number of infected farms in the seed farm's community I_C with the total number of infected farms I.

One way to interpret our results is to use a receiver operating characteristic (ROC) curve. A ROC curve is a plot of the true-positive-rate (TPR) against the false-positive-rate for a binary classifier. In our case, our community structure makes a prediction that any farm in the seed farm's community C will become infected (test positive) and any farm not in C will not become infected (test negative).

We run an epidemic and observe the results. The number of infected farms in C is the number of true-positives (TP), the farms in C that have not been infected are false-positives (FP), the true-negatives (TN) are all the farms not in C while the false-negatives are the infected farms not in C. In the notation introduced above

these classifications are given by

$$\begin{aligned} \text{TP} &= I_C && (4.32)\\ \text{FP} &= S_C - I_C && (4.33)\\ \text{TN} &= (N - S_C) - (I - I_C) && (4.34)\\ \text{FN} &= I - I_C. && (4.35) \end{aligned}$$

The true-positive-rate (TPR) also known as the sensitivity is

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} = \frac{I_C}{I} \tag{4.36}$$

and the false-positive-rate (FRR) is given by 1−the true-negative-rate (1−the specificity)

$$\text{FPR} = 1 - \frac{\text{TN}}{\text{FP} + \text{TN}} = \frac{S_C - I_C}{N - I}. \tag{4.37}$$

With S_C the size of community C. Tests resulting in a high TPR and low FPR will appear on the top left of the ROC plot and are indicative of an accurate test. If a point appears below the line from (0,0) to (1,1) then flipping the tests prediction would be a better test.

We can directly compare different partitions by considering the seed farm's community in the different partitions and computing the associated ROC score.

4.4.1 UK simulation results

As well as examining the predictions of the various partitions found by the Newman-Girvan and Spatial null models we also use three intuitive partitions of the UK; the 68 individual counties, the six regions of Scotland, North England, the Midlands, Wales, South-West England and East England, and the three nations of Great Britain; England, Scotland and Wales. We also examine one additional partition, an amalgamation of the farm-level NG partition with the inclusion of all the farms in the source farms county.

In chapter 3 we compared simulation results between different networks. To do this we used epidemic size as our measure and obtained an average epidemic size by averaging across many simulation runs. We were then able to obtain a confidence in our results by bootstrapping. Here we are more focused on which farms become infected rather than the average numbers infected. Hence, we analyse the simulations from the perspective of a farms risk of infection. [iv]

[iv] A similar methodology was employed by Tildesley et al. [2008].

For each seed county we have 5,000 simulations. We record the number of times a farm becomes infected during the first x weeks of an epidemic. We can calculate TPR and FPR for the collection of farms that have been infected in the first x weeks of at least $p\%$ of simulations.

The community which acts as our binary classifier is the community that most of the farms from the county in question are members of. That is, if we are seeding epidemics in Cumbria and examining the Counties partition we would chose the whole county of Cumbria as the test community. For non-county based partitions such as the Farm NG partition we determine which of the communities proposed by the partition farms in Cumbria fall into and then set that community as our test community. The communities considered for each of the three counties are shown in figures 4.25 - figures 4.27.

We can also look at how TPR varies with p and x. The parameters p and x are threshold parameters, with p representing the minimum risk a farm has of becoming infected and x showing how that risk increases throughout the epidemic.

Let us begin by analysing the results for Cumbria and focus on the results for simulations with $\beta = 1$ and $T = 21$. For each of the partitions considered we obtain a point on the ROC graph (figure 4.28 A) by calculating the true positive rate and false positive rate for a partition for the set of farms that have been infected within the first 4 weeks in at least 2% of simulations. As figure 4.28 A) shows partitions with larger communities such as those found using the spatial null-model and the nations of GB have very high false positive rates whilst partitions with many communities have lower false positive rates, the prime example being the counties partition.

The Nations partition yields a high TPR=0.93. Cumbria is in the community with the rest of England; this is a very large community and gives rise to a very high FPR=0.67. At the other end of the spectrum is the Counties partition. Only the farms in Cumbria are predicted to become infected. This leads to a low FPR=0.05 but only successfully identifies 0.16 of the infected farms. In the middle of these two extremes lies the Regions partition with a TPR=0.87 and a FPR=0.22. The Regions partition offers a decent TPR but the FPR result would correspond to monitoring/ controlling over 15,000 farms. Next we look to the county-level partitions found from the community detection algorithms to see if we can improve TPR or at least reduce FPR. The County NG partition gives a similar result to the Regions partition with a TPR=0.87 - the same as the Regions TPR but with an improved FPR=0.20 corresponding to 1,600 less farms being wrongly classified. Moving to the County SPA partition we retrieve our highest TPR=0.94 drastically out performing the

Figure 4.25: Plot of the test community from the Farm-County NG partition for Cumbria. The farms in the test community are marked in grey, farms that are infected at $p = 2\%$ are marked with small dots, those infected at $p = 4\%$ are marked with large dots, and those infected at $p = 10\%$ are large dots with a black centre. Farms that were infected at these p values and are members of the community are coloured orange while those that are outside the community are coloured green. The network diagram of those same farms is shown on the right.

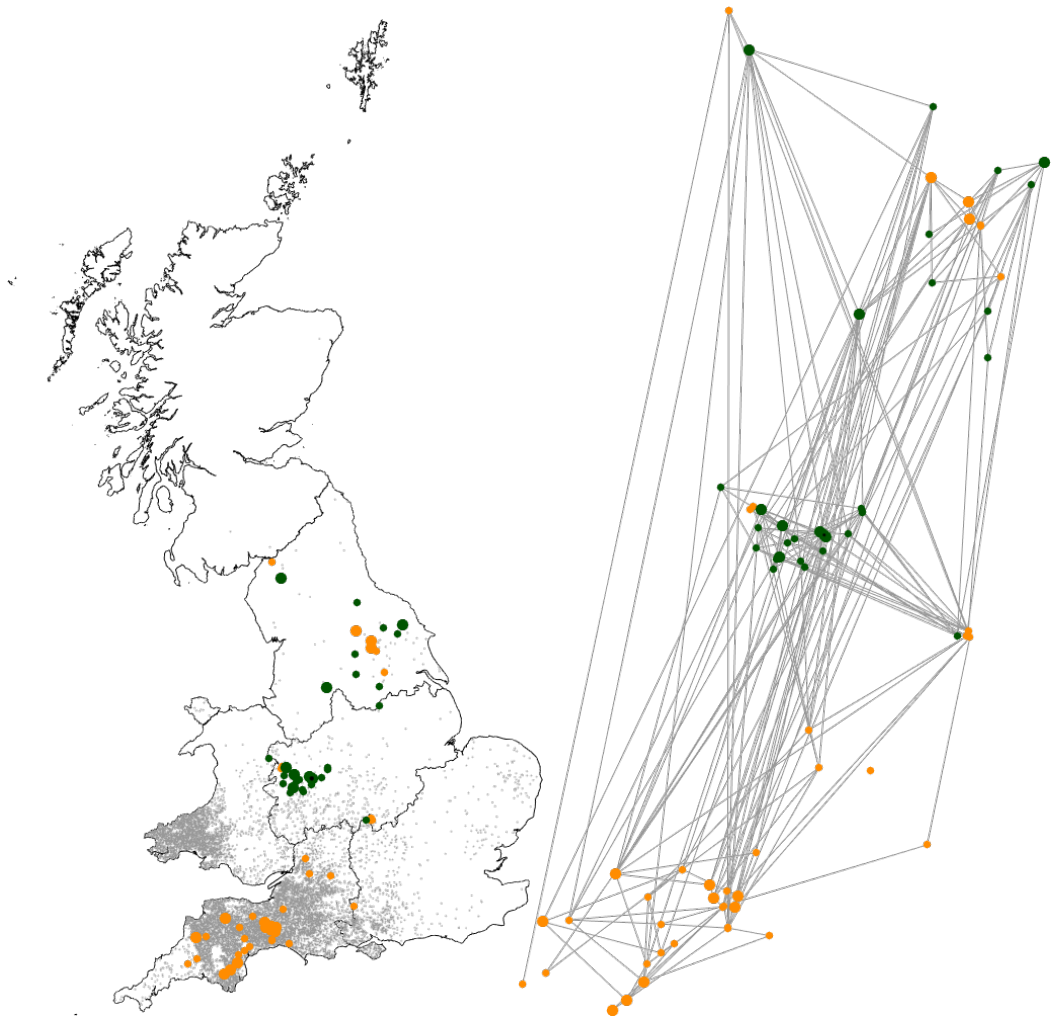

Figure 4.26: Plot of the test community from the Farm-County NG partition for Devon. The farms in the test community are marked in grey, farms that are infected at $p = 2\%$ are marked with small dots, those infected at $p = 4\%$ are marked with large dots, and those infected at $p = 10\%$ are large dots with a black centre. Farms that were infected at these p values and are members of the community are coloured orange while those that are outside the community are coloured green. The network diagram of those same farms is shown on the right.

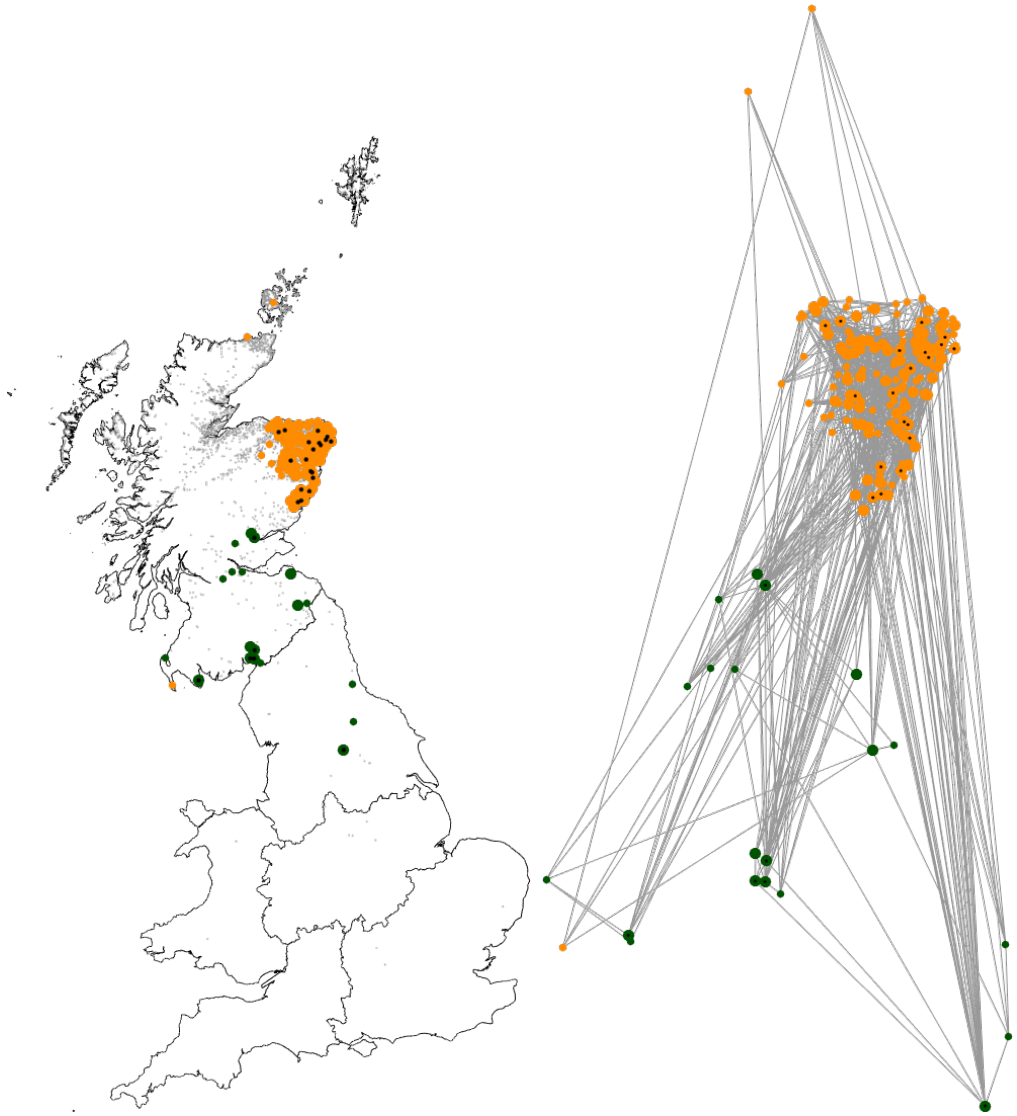

Figure 4.27: Plot of the test community from the Farm-County NG partition for Aberdeen. The farms in the test community are marked in grey, farms that are infected at $p = 2\%$ are marked with small dots, those infected at $p = 4\%$ are marked with large dots, and those infected at $p = 10\%$ are large dots with a black centre. Farms that were infected at these p values and are members of the community are coloured orange while those that are outside the community are coloured green. The network diagram of those same farms is shown on the right.

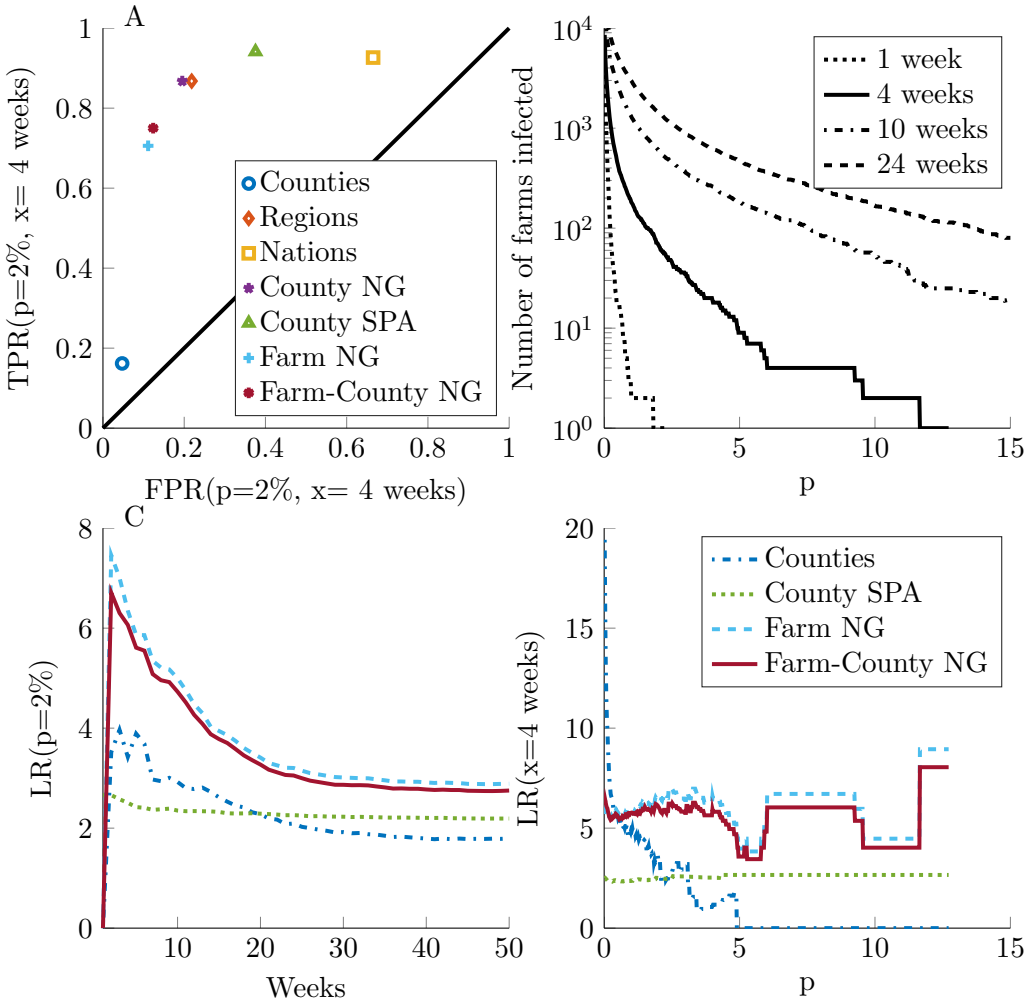

Figure 4.28: Epidemic predictions for outbreaks seeded in Cumbria. The disease parameters for these simulations were: $\beta = 1$ and $T = 21$ days. A) ROC plot for the set of farms that have become infected in the first 4 weeks of at least 2% of simulations using different partitions as predictors. B) The total number of farms that have been infected in the first 1, 4, 10 and 24 weeks of at least p% of simulations. C) The TPR for farms infected within the first x weeks in at least 2% of simulations. D) TPR after 4 weeks as a function of the percentage of simulations farms have been infected in.

Nations partition with an FPR=0.38. At the farm-level using the NG null-model we find a FPR=0.11 and a TPR =0.71. We consider also a modification of this partition and define the source farms community to be that found by the NG model but also to include the farms in the source farms county, calling this the Farm-County NG partition. This inclusion increases the size of the source farms community, increasing FPR but may capture more local connections and hence increase the TPR. For this new farm-county partition the TPR=0.75 which is an increase over the Farm NG partition with only a modest increase in FPR to 0.12.

Obviously we could examine the TPR and FPR results for a range of x and p values. Focusing on $x = 4$ weeks is a prudent choice however as if a livestock disease outbreak was to occur within the UK it is reasonable to assume that it would take some time to be detected, and once it had been detected control measures such as movement ban would be put in place.

4.4.2 Sensitivity to threshold parameters

Figure 4.28 B shows how I changes with the threshold parameters x and p. For small values of p a farm only has to be infected in a few simulations to contribute to I. A high values of x gives long period for a farm to become infected and contribute to I. Hence, low p and high x values yield the highest I values. Within the first $x = 4$ weeks, I =8,547 farms were infected in at least one simulation ($p = 0.0002\%$). This number quickly drops as we increase p to try to capture the more at risk farms. At $p = 1\%$, $I = 182$ farms dropping to 68 farms when $p = 2\%$. The number of farms consistently infected within the first week of the epidemics is very small, only 2 farms are infected across at least 1% of simulations.

The size of I will determine the TPR and FPR values with I having a more pronounced effect on TPR than FPR. For small numbers of infected farms the false-positive-rate is essentially a ratio of the number of farms we expected to become infected (the community size, S_C) to the total number of farms, N. To allow us to compare the effects of changing x and p on different partitions we use the ratio of TPR to FPR. This ratio is known as the positive likelihood ratio (LR_+).

The LR_+ ratio with $p = 2\%$ fixed and x being allowed to vary is shown in figure 4.28 C) for the Counties, County SPA, Farm NG, and Farm-County NG partitions. At $x = 1$ only one farm is consistently infected, $I(p = 2\%) = 1$. This farm is only picked up by the County SPA partition with the other partitions having LR_+=0. For the County NG and Farm-County NG partitions a peak in LR_+ is seen at $x = 2$ which begins to decay sharply before levelling off at around $x =$ 25 weeks. There is little difference in performance between these two partitions

though the initial peak for the County NG partition is slightly higher than that for the Farm-County NG partition. Both of these partitions however outperform the Counties and County SPA partitions; with the former being penalised for being too localised giving a low TPR, and the latter from being too general giving a high FPR.

When we fix $x = 4$ weeks and vary p we once again find vary little difference between the Farm NG and Farm-County NG partitions (figure 4.28 D)). For both partitions the LR_+ remains at around 6 for $p < 5\%$. For $p > 5$ the number of infected farms is small (figure 4.28 A)); as such not much weight can be given to results for higher values of p.

The Counties and County SPA show differing behaviour. For the Counties partition low p results in high LR_+ values. Epidemics are seeded in Cumbria meaning we are guaranteed that at least one farm will be in the Counties community in every simulation. As p increases the performances of the Counties partition deteriorates as the epidemics spread beyond Cumbria. Cumbria is placed in the largest community in the County SPA partition. This community is so large that TPR and FPR remain constantly high regardless of p and thus the LR_+ remains small and constant.

4.4.3 Spatial heterogeneities

The heterogeneous nature of the UK cattle network does lead to varying levels of prediction accuracy for the different counties examined. A summary of the TPR and FPR values for all three counties is given in table 4.3 with $p = 2\%$ and $x = 4$ weeks.

For Devon (figure 4.29) the Nations and Counties partitions once again appear at the extremes of the ROC plot. However, in contrast to Cumbria, we now see a much tighter clustering of the other partitions. The Farm-County NG partition has the largest TPR=0.59 but with a slightly higher FPR than the Farm-Level NG partition. These two partitions outperform the Regions, County NG and County SPA partitions.

For Aberdeen all the TPR values are high but there is a redundancy in some of the partitions. The County NG, Nations and Regions partitions all place farms in Aberdeen in the same community and hence they all receive the same ROC score (figure 4.30 A). Outbreaks are more localised in Aberdeen than in Cumbria and Devon with the Counties only partition performing well, outscoring the farm-level NG. The Farm-County NG partition does offer an improvement on both those partitions with a TPR of 0.90. The redundant partitions have high TPRs=0.98 which

	Cumbria		Devon		Aberdeen	
Partition	TPR	FPR	TPR	FPR	TPR	FPR
Counties	0.1618	0.0472	0.2267	0.0631	0.8653	0.0322
Regions	0.8676	0.2184	0.4133	0.1931	0.9845	0.1687
Nations	0.9265	0.6651	0.9867	0.6650	0.9845	0.1687
County NG	0.8676	0.1953	0.4667	0.2177	0.9845	0.1687
County Spa	0.9412	0.3755	0.4667	0.2428	1.000	0.3743
Farm NG	0.7059	0.1112	0.5333	0.1751	0.8238	0.0716
Farm-County NG	0.7500	0.1237	0.5867	0.1996	0.8964	0.0751

Table 4.3: TPR and FPR for different partitions for epidemics seeded in Cumbria, Devon and Aberdeen for farms infected the first 4 weeks of at least 2% of simulations with $\beta = 1$ and $T = 21$ days. The first three partitions are defined geographically, the next three have been found using community detection algorithms while the final partition is an amalgamation of the Counties partition with the farm-level NG partition.

is close to the TPR of the County SPA partition with a TPR=1.00. However, the County SPA partition has a much higher FPR as it includes counties from the North of England (figure 4.14).

A summary of the TPR and FPR values for each of the partitions considered in each county is given in table 4.3.

Maps (figures 4.25-4.27) reveal the spatial differences between the farms most frequently infected by epidemics seeded in each of the counties. These results mirror those seen in chapter 3 but give a more detailed picture as to why community detection is more successful for some counties than for others.

For all three counties there exist farms that are frequently infected that were not placed into the predicted community. These farms have high in-degrees, and are therefore very well connected within the network. This is most notable for Devon (figure 4.26) where a large cluster of farms in Herefordshire are frequently infected but were not placed in the same community as the farms from Devon.

4.4.4 Sensitivity to disease parameters

As mentioned above simulations were carried out for $\beta \in \{1, 2, 3, 4, 5\}$ and $T \in \{7, 14, 21, 28\}$ days. The effect of changing these parameters on epidemic size and and on the likelihood ratio are shown in figure 4.31. We fix x at 4 weeks and compare results with the threshold parameter $p = 2\%$.

The transmission parameter has a larger impact on the number of farms infected and on the likelihood ratio than the infectious period. Results for $\beta = 1$

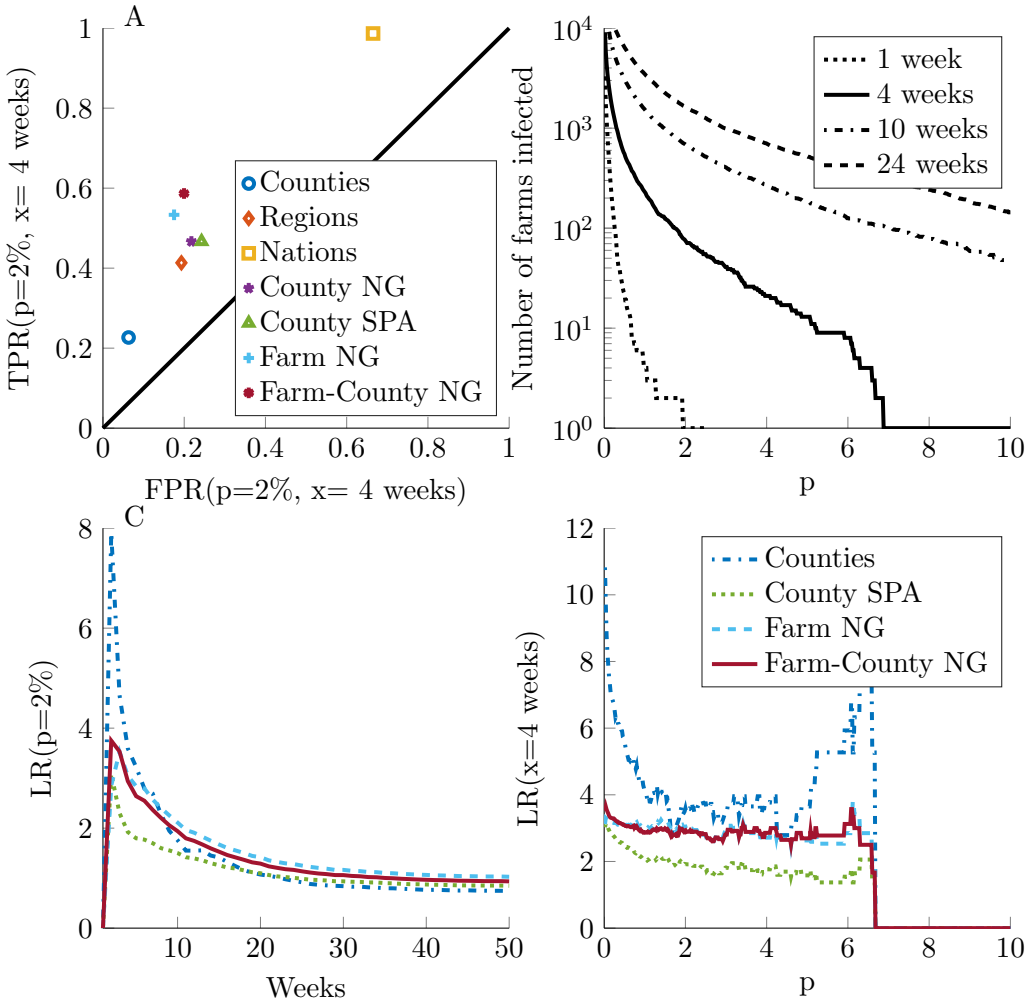

Figure 4.29: Epidemic predictions for outbreaks seeded in Devon. The disease parameters for these simulations were: $\beta = 1$ and $T = 21$ days. A) ROC plot for the set of farms that have become infected in the first 4 weeks of at least 2% of simulations using different partitions as predictors. B) The total number of farms that have been infected in the first 1, 4, 10 and 24 weeks of at least p% of simulations. C) The TPR for farms infected within the first x weeks in at least 2% of simulations. D) TPR after 4 weeks as a function of the percentage of simulations farms have been infected in.

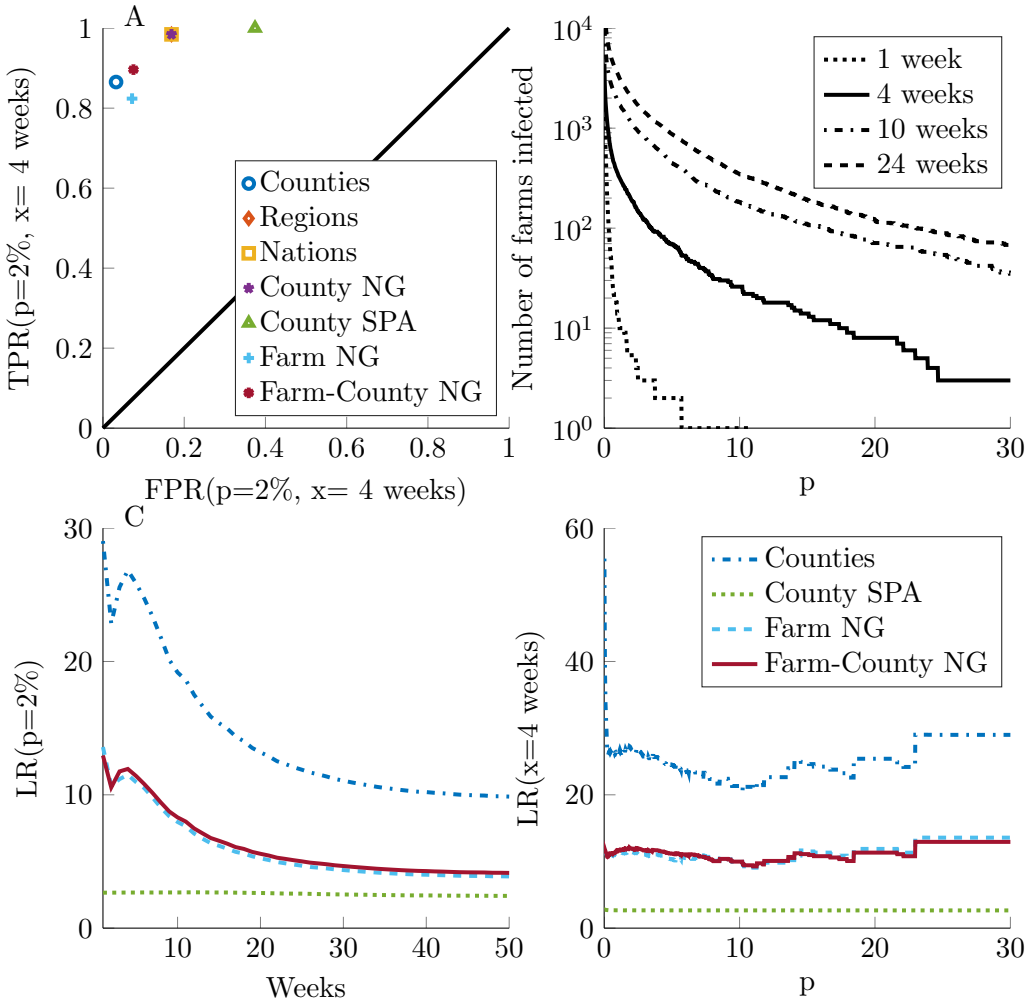

Figure 4.30: Epidemic predictions for outbreaks seeded in Aberdeen. The disease parameters for these simulations were: $\beta = 1$ and $T = 21$ days. A) ROC plot for the set of farms that have become infected in the first 4 weeks of at least 2% of simulations using different partitions as predictors. B) The total number of farms that have been infected in the first 1, 4, 10 and 24 weeks of at least p% of simulations. C) The TPR for farms infected within the first x weeks in at least 2% of simulations. D) TPR after 4 weeks as a function of the percentage of simulations farms have been infected in.

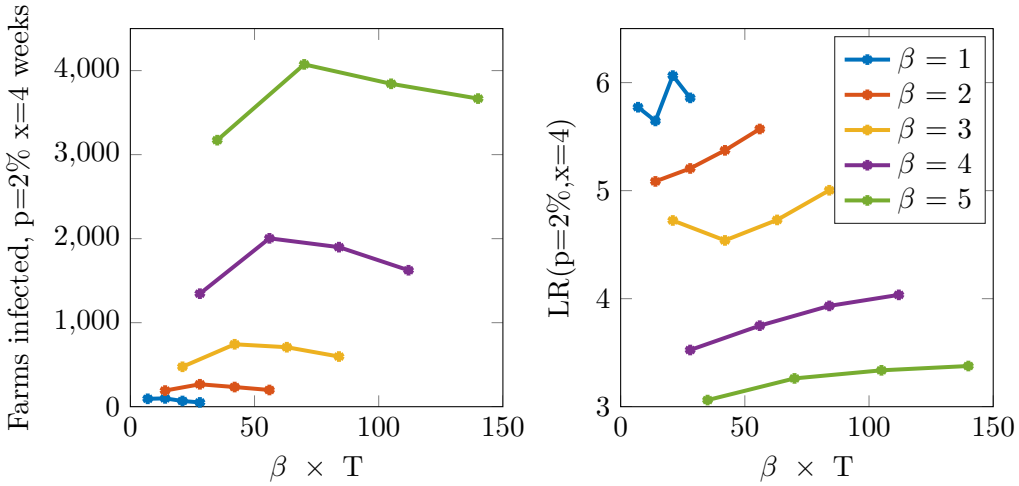

Figure 4.31: The effect of varying disease parameters on epidemic size (left) and positive likelihood ratio (right). All analysis takes place 4 weeks into the epidemic or $\beta \in \{1, 2, 3, 4, 5\}$ per day and $T \in \{7, 14, 21, 28\}$ days. We analyse simulations seeded in Cumbria examining farms that have been infected in at least 2% of simulation.

vary between $LR_+ = 5.6$ and 6 but drop to between 3 and 3.4 for $\beta = 5$. This is not surprising as the high β increases the probability of transmission along all edges making it more likely that edges with a low weighting could transmit disease.

A more subtle effect is seen when transmission is kept constant and the infectious period T is varied. For all values of β we observe that increasing T can actually lead to smaller epidemics. In fact we observe an increase from $T = 7$ days to $T = 14$ days and then a decrease through to $T = 28$ days. This appears contradictory, increasing T directly increases R_0 and should produce larger epidemics. The final sizes of the epidemics we have simulated are indeed larger for higher values of T. Low numbers of infected farms early on in the outbreak are explained by the take-off limit we imposed on simulations.

Imagine we run two simulations, one with a high infectious period T_H and one with a lower infectious period T_L. We seed each simulation in the same farm, F. We assume F is weakly connected to the network, i.e. has a low centrality, and therefore has a low probability of causing an outbreak. For the simulation seeded with the short infectious period, F will remain infected for T_L days. If, as expected, F does not infect any other farms then the take-off limit will not have been reached and a new seed farm will be chosen at random. In this way outbreaks seeded with a low infectious period have to reach the take-off threshold quickly in order to be counted so we are preferentially choosing outbreaks that take-off more rapidly.

On the other, if we seed F using the high infectious period then F will remain

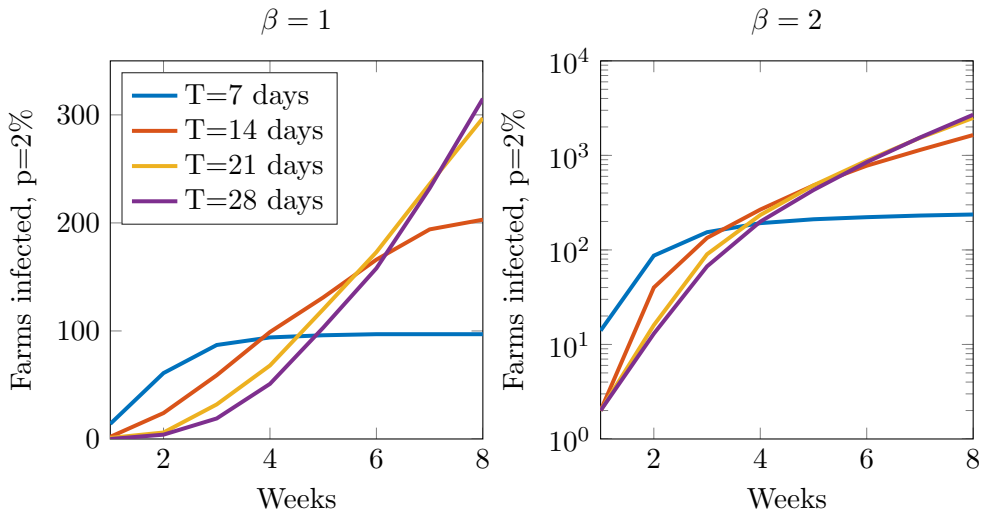

Figure 4.32: Plot showing the affect that varying the infectious period T has on the number of infected farms in the early stages of outbreaks when a take-off limit is imposed. On the left we have the transmission parameter $\beta = 1$ and on the right $\beta = 2$. Simulations are seeded in Cumbria and we analyse the farms infected in at least 2% of outbreaks.

infectious for T_H giving a greater chance of transmitting infection and leading to an outbreak. The longer infectious period allows the outbreak to remain small for much longer before reaching the take-off limit.

Ultimately, the outbreak with the longer infectious period will be larger. Figure 4.32 demonstrates that outbreaks with shorter infectious periods can have larger sizes early in the epidemic. In our case we look at outbreaks with infectious periods of 7,14,21, and 28 days and make our analysis 4 weeks into the outbreak. By this time outbreaks with $T = 14$ days have caught up with those having a $T = 7$ days, while those with higher infectious periods are still in their early stages.

Similar results are seen for epidemics seeded in Devon and Aberdeen and are given in the appendix C.

4.5 Discussion

We have used the community detection method to analyse the UK and Turkey cattle movement networks in a novel manner and applied the resultant partitions for the UK to epidemic predictions.

The only previous application of community detection to networks generated from between farm movement of livestock was carried out on the network of sheep

movements by Kao et al. [2006] and continued by the same authors in Green et al. [2006]. Their work focused on farm level networks which were aggregated over four week periods for sheep movements throughout 2003.

There are some similarities in the maps generated by Kao et al. [2006] of the sheep communities and our map of the farm level cattle communities, figure 4.10. Scotland and the North of England both appear as distinct communities in both sets of analyses. For the sheep network the South of England and South of Wales appear as one large community whereas for the cattle network they are more disjoint.

We have shown on the county level that the cattle communities of the UK do change with time but that the North of the UK appears robust to this change. It would be interesting to include coupling for the farm level UK networks so as to track the temporal evolution of the community structure at a finer scale. This would be computationally expensive as having M time point would increase the size of the adjacency matrix from $[N \times N]$ to $[(MN) \times (MN)]$.

In searching for partitions of a network using community detection methods we are ultimately searching for new ways to classify nodes. Finding these hidden classifications can prompt discussion as to what properties nodes share that might cause them to be placed in the same community. In the networks we have studied geographical distance seems to be the driving factor behind community assignment. This is more evident in the Turkish cattle network where the divide between different communities quite often falls close to provincial boundaries. This is presumably due to restrictions on trade between provinces in Turkey.

The UK did not show geographical bias to the same extent as Turkey; except for the notable exception of Scotland. In almost all algorithms we implemented the farms and counties of Scotland were placed in the same community. The exception to this was our implementation of the spatial null-model.

The result of the spatial null model was to bring much of the North of England into the Scottish community and to create a non-contiguous community between the South West of England and East Anglia. Due to computational limitations we could only implement this algorithm at a county-level. Given the resources it could prove useful to extend this method to the full farm level network for both the UK and Turkey. This may yield a more resolved community structure which reveals deeper insight into the formation of trade links between farms than proximity alone.

The ROC plots reveal that it is possible to have a high level of predictive accuracy using community structure but this comes at a cost of many false positive which would consume resources in the event of an outbreak. The combining of the Farm-Level NG partition with the Counties partition improved results for all

counties considered.

Whilst the results of the epidemic simulations for the UK using community detection as a predictor were not startling they do lend themselves to some general truths. Real world networks are highly heterogeneous particularly if they are embedded spatially. The cattle network in the UK behaves differently in different regions and therefore a control/ surveillance strategy that appears suited to one region may fail significantly when applied to others.

We have also not at this point considered markets. As demonstrated in chapter 3 markets can dramatically alter the dynamics of outbreaks, however it is still not known how best to model transmission at a market. It is also unclear how markets should be treated in a community detection context. Should they be treated as a separate type of node, or be directly incorporated into the network?

Similarly, knowing the production type of each farm (beef or dairy) and incorporating this into the community detection algorithms may provide deeper insights into the structure of the cattle network. As some farms have mixed animal types it may even be possible to overlay the sheep and pig movement data to create a multilayer network with different types of nodes linked through different types of edges. Recently work has been done by Kivelä et al. [2014] to lay a foundation for analysis of this type of network. Applying this analysis to the whole livestock network for a country would require amalgamating many data sources.

Parameter sensitivity suggests that predictions can be more accurate for a disease with a low transmission rate but a longer infectious period. It is possible then that this work could be applied to a slow-burning disease such as bovine tuberculosis (BTB). As it stands, our farm based SIR model would be unable to capture the dynamics of BTB, with an individual based model incorporating a latent period being more appropriate [Brooks-Pollock et al., 2014; O'Hare et al., 2014].

Chapter 5

Modelling FMD in Turkey

While we could embark directly on the development of a national scale model assuming that the spread and virulence of FMD in Turkey will be similar to that observed during the 2001 outbreak in the UK, this approach will probably prove flawed. As discussed in previous sections of this thesis farming practices differ between the two countries as do attitudes towards FMD and, as a result the available control options in Turkey are very different to those used in the 2001 outbreak. Bearing this in mind we thought it prudent to examine specific Turkish farms which have had multiple outbreaks of FMD in order to derive parameter distribution specific to Turkey.

Farm 8149 is the village of Yurtbasi located in the Merkez district of the Elazig province. It has had nine outbreaks of FMD between May 2002 and May 2012, with seven of those occurring after July 2007. Of these seven, there has been one case of type A, four of type O and most recently two of type Asia-1. The population of the farm has ranged from 3600 to 6000 cattle and was comprised of 785 holdings[i].

Using an individual cattle model in a specific farm should allow us to gain an understanding of the epidemiology of the disease as well as the effects of vaccination and waning immunity. It may be the case, however, that parameter values inferred for one farm are completely different for a farm with differing demography. To check for this we chose a second farm - farm 25,372 is the village of Kalecik in the Kozakli district of Neveshir. This farm is much smaller with cattle population ranging from between 100 and 650 cattle coming from 118 holdings. As with farm 8,149 farm 25,372 was chosen because of its history of multiple outbreaks of FMD since 2007. The locations of both farms are shown in figure 5.1.

In this chapter we construct a model for both farms. We then parametrise the

[i]Though not all of these holdings are active every year

Figure 5.1: Locations of the farms mentioned in the text. Farm 8159 (yellow) is the village of Yurtbasi located in the Merkez district of the Elazig province. Farm 25,372 (red) is the village of Kalecik in the Kozakli district of Neveshir province. All 81 provincial outlines are shown. The satellite image is courtesy of Google Maps.

model using movement and outbreak data from both farms in a Bayesian framework. Simulations are then performed using posterior parameter distributions.

5.1 The model

In order to simulate FMD within a farm we construct a basic SEIRS compartmental model. This allows for a latent period and waning immunity. The model is refined by including classes for vaccination V, and maternally derived immunity M. The proportion of cattle in each class will be denoted by lower case letters. The daily transitions between the different classes will be determined by Poission processes. The rates for these processes (save for the S→E transition) will be inversely proportional to the average time an animal is expected to stay in that class.

Not governed by an individual rate is the transition from susceptible to infected (S→E), instead this rate is proportional to the proportion of infectious cattle on the farm, i. Susceptible cattle therefore become infected with probability

$$\mathbb{P}(\mathrm{S} \to \mathrm{E}) = 1 - \mathrm{e}^{-\beta i} \tag{5.1}$$

where β is the transmission rate.

The E, I, R, V and M classes will each be composed of multiple states[ii]. Allowing for multiple states gives some control over the distribution of the period spent in a class. This period moves from being exponentially distributed when there is only one state to approaching a step-function when there are many states.

[ii]In the literature the terms states, classes and compartments are used interchangeably by different authors

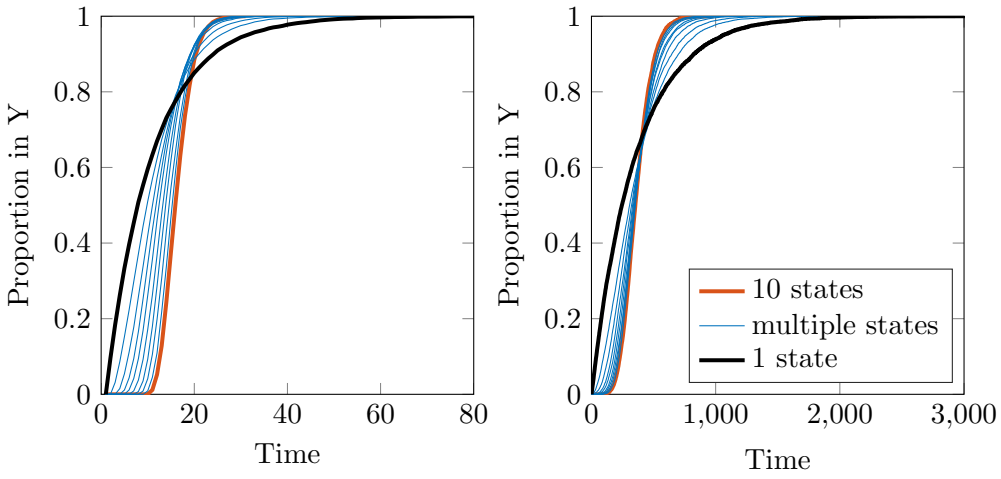

Figure 5.2: Figure indicating the effect having multiple states has on the time it takes for transitions between class X and class Y. On the left hand side the period is 10 days while on the right hand side the period is 360 days.

[Keeling and Rohani, 2008] (figure 5.2). This method allows us to guarantee that a cow will remain in one class for at least as many days as there are states in the class. This also incorporates a time delay into the dynamics while still keeping them Markovian. As the number of states becomes much greater than the average period divided by the time-step the probability of moving between states approaches one ensuring the cattle will move continuously through states recovering the dynamics we imposed for the I→R transitions in the farm level model of chapters 3 and 4 which were dictated by a fixed recovery period.

Cattle moving to the farm may be placed into the vaccinated category. Each cow moving onto the farm will be vaccinated with probability $p = v_c$, where v_c is the vaccine coverage. Vaccinated cattle will be uniformly distributed among the vaccinated states. Those cattle not vaccinated will be susceptible.

Turkey has ongoing vaccination campaigns that run at least twice a year. As the exact dates of these campaigns were not known to us we implement them twice yearly at the end of March and September each year. This corresponds to the most common time when vaccinations were carried out in 2012 [Knight-Jones, 2014]. An attempt is made to vaccinate all susceptible cattle on the farm. Susceptible cattle are successfully vaccinated with probability $p = v_e$, with v_e the vaccine efficacy.

Further vaccination is also carried out in the event of an outbreak. In the same way as the routine vaccination an attempt is made to vaccinate all susceptible cattle with the vaccine efficacy determining which cattle become vaccinated.

As we have information on births we can incorporate waning maternal im-

parameter	symbol	prior
transmission rate	β	$\mathcal{U}(0, 10)$
number infected	n	set to 1
length of E period	σ^{-1}	Weibull(3.974,1.782) [Mardones et al., 2010]
length of I period	γ^{-1}	Gamma(3.969,1.107) [Mardones et al., 2010]
length of R period	w^{-1}	$\mathcal{U}(30, 720)$ [Pomeroy et al., 2015]
length of V period	ν^{-1}	$\mathcal{U}(30, 180)$ [Doel, 2005]
length of M period	κ^{-1}	$\mathcal{U}(1, 180)$ [Gonzales et al., 2014]
vaccine coverage	v_c	$\mathcal{U}(0, 1)$
vaccine efficacy	v_e	$\mathcal{U}(0, 1)$
boost efficacy	v_b	$\mathcal{U}(0, 1)$
time-to-control	T_C	$\mathcal{U}(14, 35)$
NSP prevalence	r_{NSP}	$\mathcal{U}(0, 1)$
Number of E states	NoE	set to 2
Number of I states	NoI	set to 2
Number of R states	NoR	random integer from 1 to 10
Number of V states	NoV	random integer from 1 to 10
Number of M states	NoM	random integer from 1 to 10

Table 5.1: Parameters in the model with the symbol used and any prior knowledge incorporated with references

.

munity into the model as class M. Each day a number of newborn cattle will be placed into M with probability $p = 1 - s$ while the remainder will become susceptible.

5.1.1 Free parameters

As described above the model has many parameters. Here we will outline each of these parameters and give prior estimates of the distributions that may describe them. Unless stated, parameters are not bounded by these prior distributions and their true values may lie outside of this range. This will be accounted for in the parameter inference schemes outlined later in this chapter. A summary of the parameters and their prior distributions is given in table 5.1.

Infection will be determined by the transmission rate β. In most cases studied a value for β is not presented with the R_0 being reported instead. At a farm level Ferguson et al. [2001b] estimated R_0 between farms in the UK to be between 3.5-4.5, which dropped to 1.6 in the presence of movement restrictions. Estimates within a herd vary; when reporting on an FMD outbreak in Bolivia Gonzales et al. [2014] found R_0 to range from 1.02 to 2.68, however this was not a true value of R_0 as some cattle in the herd were vaccinated. In a controlled study of FMD transmission

between cattle Orsel et al. [2007] found $R_0 > 1$ but placed no upper bound on the value. The same authors found that for calves (aged 8-10 weeks) $R_0 = 2.52$. In both studies the authors found that vaccination drove R_0 below the threshold value of 1.

As we incorporate vaccination into our model and are not challenging a fully susceptible population with FMD we would expect our estimate of R_0 to be in the range given by Gonzales et al. [2014]. Allowing for an infectious period of at least 2 days would suggest that $\beta = R_0 \times \gamma^{-1} = 2.68/2 \approx 1$. This is only a rough estimate and we will allow for an unusually high value of β by selecting β uniformly from 0 to 10. [iii]

Outbreaks will occur at the dates given from the data. The number of cattle infected n when an outbreak is seeded in the farm will influence the probability of the outbreak taking-off. However, we will reduce the parameter space of the model by setting $n = 1$.

As discussed in the introduction there is some debate over the duration of the exposed and infectious periods for FMD in cattle. We will opt to use the distributions proposed in the meta-analysis put forward by Mardones et al. [2010] pertaining to the O serotype of FMD. They offered an exposed (latent in their terminology) period given by a Weibull distribution with scale parameter $a_W = 3.974$ and shape parameter $b_W = 1.782$. For the infectious period we use a Gamma distribution with shape parameter $a_\Gamma = 3.969$ and scale parameter $b_\Gamma = 1.107$.

Estimates for the period for the R, V, and M classes are hard to find with few longitudinal studies having been carried out. What information is available is discussed in the introduction chapter. With the broad consensus being that vaccination and maternally derived immunity (if any exists) should last no more than 6 months we set prior distributions of $\mathcal{U}(30, 180)$ days and $\mathcal{U}(1, 180)$ days for the V and M periods respectively. We will allow for a long recovery period, which, may be life-long [Pomeroy et al., 2015] and use $\mathcal{U}(30, 720)$ days as the prior distribution for the R class.

While the date the outbreak was confirmed is assumed to be known, the date of the reactive vaccination program is not. To account for this a time-to-control parameter T_C is included in the model which represents the delay from the start of the outbreak to the introduction of the vaccination programme. We use $\mathcal{U}(14, 35)$ days as the prior distribution for T_C as we would expect some delay in the implementation of the vaccination but that it would be carried out within the duration of the outbreak

As the E and I periods are likely to be less than 4 or 6 days we will set

[iii] We will denote $\mathcal{U}(a, b)$ as the uniform distribution ranging from a to b.

the number of states in these classes to 2. For the R,V, and M classes the number of states will be drawn uniformly from the integers 1 to 10. This was deemed appropriate as increasing the number of states beyond ten has little effect on the overall dynamics but does increase computation time.

5.1.2 Model summary

- Individual (cattle) level model.
- Cattle are born, die and move on and off the farm according to the (cleaned) TurkVet data.
- The disease model is SEIR.
- Each cow has a disease class and state.
- Cattle moving onto to the farm are vaccinated with probability v_c otherwise they are susceptible.
- Susceptible cattle on the farm may be vaccinated at certain dates with probability v_e.
- On vaccination dates cattle already vaccinated may be boosted back to the first vaccination state with probability v_b.
- Cattle born onto the farm are given maternally derived immunity with probability $1 - s$.
- All forms of immunity wane and eventually result in the cattle becoming susceptible.
- Outbreaks occur on dates given by the outbreak data.
- One susceptible cow is randomly infected on the first day of the outbreak.
- During an outbreak births, deaths and movements off the farm continue as usual.
- During an outbreak movement to the farm are placed into a holding pen where they cannot become infected.
- On the day after the outbreak ends cattle in the holding pen are released into the main farm.
- All updates take place with a daily time-step.

5.1.3 Model implementation

The model is initialised with a set of parameters. We generate a transition matrix which determines which classes will transition to which other classes. The number of states within each class is directly incorporated into this matrix. The model treats classes and states equivalently allowing us to remove the distinction and refer to both as states.

Except for the S→E transition the probability of moving between two states X_k to X_{k+1} where class X has N_X states is

$$\mathbb{P}(X_k \to X_{k+1}) = 1 - \mathrm{e}^{-\frac{N_X}{T_X}} \tag{5.2}$$

which can be stored in memory. The probability of moving from the last state in X to the first state in Y is also determined by the number of states in X.

A state matrix ($M \times N_{\text{Total}}$) is created which keeps track of all the cattle that will appear during the simulation and has a column for each of the N_{Total} states. Those cattle that start on the farm are marked as susceptible in the state matrix. A proportion, v_c of these cattle are selected to be placed into V_1. A proportion, r_{NSP} are also placed into R_1. If the same cow was selected to be made both vaccinated and recovered then recovery took precedence.

Now the simulation begins progressing with daily time-steps. At the start of each day the proportion of susceptible animals on the farm, s is calculated. Next the movements to and from the farm as well as births and deaths take place. As described above, a proportion of the movements to the farm will be vaccinated and a proportion of the births will be given maternally derived immunity, with the remaining cattle being susceptible.

The probability of each cow transitioning to the next state is then calculated and any updates are performed. If the time-step date equals a routine vaccination date then vaccination will be applied to the farm.

If the date is equal to an outbreak date then one susceptible cow is chosen to be infected. Now the probability of the S→E transition which depends on i will have to be calculated. If the time-step date equals an outbreak start date then one susceptible cow is made infectious.

During the outbreak, as defined by the data, cattle that were born or moved onto the farm were placed in a separate holding-pen. The farm is made up of several holdings though we have no information as to which holdings were infected. It is assumed that the holdings that were unaffected by the outbreak will keep their cattle separate from infected cattle. The holding-pen allows trade to continue on the

farm without running the risk of introducing new susceptible cattle to the outbreak. When the outbreak is over the farms in the holding-pen are moved to the main farm.

5.2 Parameter inference

Bayesian inference allows for the use of prior knowledge of parameters and therefore allow modellers to directly incorporate the beliefs of experts such as policy makers, physicians, and veterinarians. One such class of Bayesian inference schemes, Markov Chain Monte Carlo (MCMC) methods have been widely used in recent years to fit epidemiological models [McKinley et al., 2014; Neal and Roberts, 2004; Tanaka et al., 2006; Zhang and Woolhouse, 2011].

MCMC methods treat the parameters, θ as random variables that come from some distribution given the observed data x, known as the posterior distribution $f(\theta|x)$. Our belief over the true value of θ is given by the prior distribution $\pi(\theta)$. In many cases we have no informed belief in the prior in which case it is common to use default fat or objective priors.

Under-pinning MCMC methods is Bayes rule which establishes a relationship between the prior and the posterior, such that the posterior is proportional to the prior times the likelihood of the data given the parameters $\mathcal{L}(\theta|x)$

$$f(\theta|x) \propto \pi(\theta)\mathcal{L}(\theta|x). \tag{5.3}$$

MCMC is used in situations when this constant of proportionality is unavailable and proceeds by establishing a Markov chain with $f(\theta|x)$ as its underlying distribution. This chain can then be sampled from to give the posterior distribution. Thus an estimate of the posterior distribution can be found using only the $\pi(\theta)$ and $\mathcal{L}(\theta|x)$.

Once such MCMC method that is commonly employed is the Metropolis-Hastings algorithm [Hastings, 1970; Metropolis et al., 1953] outlined below (algorithm 3). This method starts at a preassigned point in parameter space θ_0 and proposes to jump to a new point θ' using a proposal distribution $q(\theta', \theta_0)$. This jump is made with probability

$$a_{\mathrm{MH}} = \min\left(1, \frac{\mathcal{L}(\theta'|x)}{\mathcal{L}(\theta_0|x)} \frac{q(\theta', \theta_0)}{q(\theta_0, \theta')} \frac{\pi(\theta')}{\pi(\theta_0)}\right) \tag{5.4}$$

which sets $\theta_1 = \theta'$ if the jump is made and $\theta_1 = \theta_0$ otherwise. This is continued until the Markov Chain converges. The possibility of accepting the jump even if the proposed parameters decrease the likelihood of the data allows the chain to explore parameter space and is necessary to sample effectively from multi-modal

distributions.

Algorithm 3 Metropolis-Hastings algorithm

$k = 1$;
$\theta_0 \sim \pi(\theta)$; ▷ Set θ_0 from the prior
while Chain not converged **do**
    $\theta' \sim q(\theta', \theta_{k-1})$ ▷ Propose new parameters
    $a = \min\left(1, \frac{\mathcal{L}(\theta'|x)}{\mathcal{L}(\theta_0|x)} \frac{q(\theta', \theta_0)}{q(\theta_0, \theta')} \frac{\pi(\theta')}{\pi(\theta_0)}\right)$ ▷ Calculate the acceptance probability
    if $RAND < a$ **then**
        $\theta_k = \theta'$
    else
        $\theta_k = \theta_{k-1}$
    end if
    $k = k + 1$
end while

The algorithm will eventually converge on the posterior. However, initial samples may be very far from this distribution, particularly if a poor choice of prior is used. As such a number of the initial samples must be discarded, this is known as a 'burn-in' period.

Nearby samples will be highly correlated. If (approximately) independent samples are required then it is possible to 'thin' the set of samples, generally by selecting every n^{th} sample. The extent of thinning is determined by examining auto-correlations between the samples. The burn-in period and thinning of the samples increase the computational time of the model.

Often the proposal distribution q is chosen to be symmetric, simplifying the acceptance ratio, with uniform priors simplifying matters further. What remains is the likelihood function, however for complicated models this is not generally tractable and other approaches must be explored.

Another class of methods which build on MCMC methods but do not require explicit calculation of the likelihood function are approximate Bayesian computation (ABC) methods. One such ABC method follows the same steps as the Metropolis-Hastings algorithm but rather than computing a likelihood an appropriate summary statistic $\sigma(\theta)$ is used instead. A simulation is carried out using the proposed parameters θ', generating new data $x' \sim \mathcal{M}(\theta')$ and the results compared with the data. A distance $\rho(\sigma(x'), \sigma(x))$ between the simulation results and the original data is determined. If this distance is below some tolerance ϵ then the new parameters are

kept with probability

$$a_{\text{ABC}} = \min\left(1, \frac{q(\theta', \theta_0)}{q(\theta_0, \theta')} \frac{\pi(\theta')}{\pi(\theta_0)}\right). \tag{5.5}$$

This algorithm (algorithm 4) proposed by Marjoram et al. [2003] is likelihood free and outputs the stationary distribution $f(\theta|\rho(\sigma(x'), \sigma(x)) < \epsilon)$. As with the Metropolis-Hastings algorithm a symmetric proposal distribution can be used which further simplifies calculations.

Algorithm 4 ABC algorithm

$k = 1$;
$\theta_0 \sim \pi(\theta)$; ▷ Set θ_0 from the prior
while Chain not converged **do**
    $\theta' \sim q(\theta', \theta_{k-1})$ ▷ Propose new parameters
    $x' \sim \mathcal{M}(\theta')$ ▷ Generate sample from model
    $d = \rho(\sigma(x'), \sigma(x))$ ▷ Compute distance measure
    if $d < \epsilon$ **then**
        $a = \min\left(1, \frac{q(\theta', \theta_0)}{q(\theta_0, \theta')} \frac{\pi(\theta')}{\pi(\theta_{k-1})}\right)$ ▷ Calculate the acceptance probability
        if $RAND < a$ **then**
            $\theta_k = \theta'$ ▷ Accept new parameters
        else
            $\theta_k = \theta_{k-1}$
        end if
    else
        $\theta_k = \theta_{k-1}$
    end if
    $k = k + 1$
end while

It is however the finer points of the algorithm that lead to complications. A sufficient summary statistic must be used and a suitable distance measure employed. The tolerance must also be selected: too low will lead to massive computation times too high and it will be impossible to accurately capture the posterior. As with the Metropolis-Hastings algorithm a test for convergence must also be employed. The Markov chain can again be thinned to remove correlations and a burn-in phase is also required.

To mitigate against some of these complications Sisson et al. [2009] built upon the idea of sequential Monte Carlo (SMC) methods and particle rejection control (PRC) to create an ABC-PRC algorithm. This algorithm has been used for parameter inference in models of bovine tuberculosis [Brooks-Pollock et al., 2014; O'Hare et al., 2014]. A simplified version of this algorithm was implemented by

Walker et al. [2010] in their analysis of the Hong Kong SARS outbreak of 2003. It is the algorithm of Walker et al. [2010] (algorithm 5) that we will use.

The basic premise of SMC methods is to move from the prior distribution $f(\theta)_1 = \pi(\theta)$ to the target posterior $f(\theta)_T$ through a series of intermediate distributions specified by a decreasing vector of tolerances $\vec{\epsilon} = \epsilon_1 \ldots \epsilon_T$. Initially a collection of N particles are drawn from the prior. A simulation is run for each particle. If the simulation results are within ϵ_1 of the data then the particle is kept. New particles are drawn from the prior until N particles have been accepted.

This set of particles $\{\theta_1\}$ is used as the distribution $f(\theta_2)$. Particles are sampled and perturbed from this distribution with model simulations being run. These particles are now accepted with the stricter tolerance ϵ_2 to generate a new distribution $f(\theta_3) = \{\theta_2\}$. This process is repeated until the target distribution $f(\theta_T) = \{\theta_{t-1}\}$ is found using the strictest tolerance ϵ_T.

Algorithm 5 ABC-PRC algorithm

$t = 1$;
$\epsilon = \epsilon(t)$
for $i = 1 : N$ **do** ▷ Sampling from prior
  $d = 0$;
  while $d < \epsilon$ **do**
    $\theta' \sim p(\theta)$ ▷ Propose new parameters
    $x' \sim \mathcal{M}(x|\theta')$ ▷ Generate sample from model
    $d = \rho(\sigma(x'), \sigma(x))$ ▷ Compute distance measure
  end while
  $\theta_t(i) = \theta'$ ▷ Populate the next distribution
end for
for $t = 2 : T$ **do** ▷ Decrease tolerance
  for $i = 1 : N$ **do** ▷ Sample from higher tolerance distributions
    $d = 0$;
    while $d < \epsilon$ **do**
      $\theta'' \sim \{\theta_{t-1}\}$ ▷ Sample from previous distribution
      $\theta' = \theta'' + \delta$ ▷ Perturb parameters
      $x' \sim \mathcal{M}(x|\theta')$ ▷ Generate sample from model
      $d = \rho(\sigma(x'), \sigma(x))$ ▷ Compute distance measure
    end while
    $\theta_t^i = \theta'$ ▷ Populate the next distribution
  end for
end for

The ABC-PRC algorithm has a number of advantages over the previously outlined methods. Of particular note is the fact that no test for convergence is required as the particles are uncorrelated at every stage. Using a decreasing range

of tolerances allows the algorithm to hone in on the target distribution. In practice this is much quicker than setting the desired tolerance at the start with Sisson et al. [2009] finding the ABC-PRC algorithm required 80% fewer simulations than simulating with the desired tolerance from the beginning.

5.2.1 Implementation

We directly implement algorithm 5 using the prior distributions outlined in table 5.1. We set the number of particles to 250.

Of key importance is the choice of summary statistic. In their work on the Hong Kong SARS outbreak Walker et al. had a time-series of newly infected individuals. They smoothed this time-series using a moving average. They then directly compared their model outputs with this curve using a Euclidean distance. The data we use here do not include a time series of newly reported cases. We have a collection of outbreaks and an estimate of the number of infected cattle during each outbreak.

Naïvely we could attempt to fit each outbreak separately but as we only know the final epidemic size this would result in an estimation of R_0 without a proper specification of β and the infectious period. Instead, we fit to the vector of outbreak sizes from the data $\mathbf{I}_D$; this is compared with the vector of simulated outbreak sizes $\mathbf{I}_M$:

$$\|\mathbf{I}_M^{(k)} - \mathbf{I}_D\| < \epsilon^{(k)}\mathbf{I}_D \tag{5.6}$$

where $\epsilon^{(k)}$ is the desired tolerance.

For example, if $\epsilon = 0.5$ then the difference between outbreak size of the simulated and reported data would have to be within 50% of the original outbreak size. We further enforce that this condition is met for each of the outbreaks in order for a set of parameters to be accepted.

In order to increase simulation speed a check is made at the end of each outbreak during a simulation. If the outbreak size is outside the desired tolerance then the simulation ends without simulating the remaining outbreaks, new parameters are then drawn and simulated.

As detailed in algorithm 5 at the weakest tolerance we will sample from the prior. For the remaining tolerances we draw new samples from the distributions generated at the previous tolerances. These samples are then perturbed before running simulations and testing summary statistics. We perturb these parameters by an amount $\delta \sim \mathcal{N}(0, \sigma^2)$. Different zero-centred Gaussian distributions are used for different parameters as the ranges the parameters are exploring vary. Many of

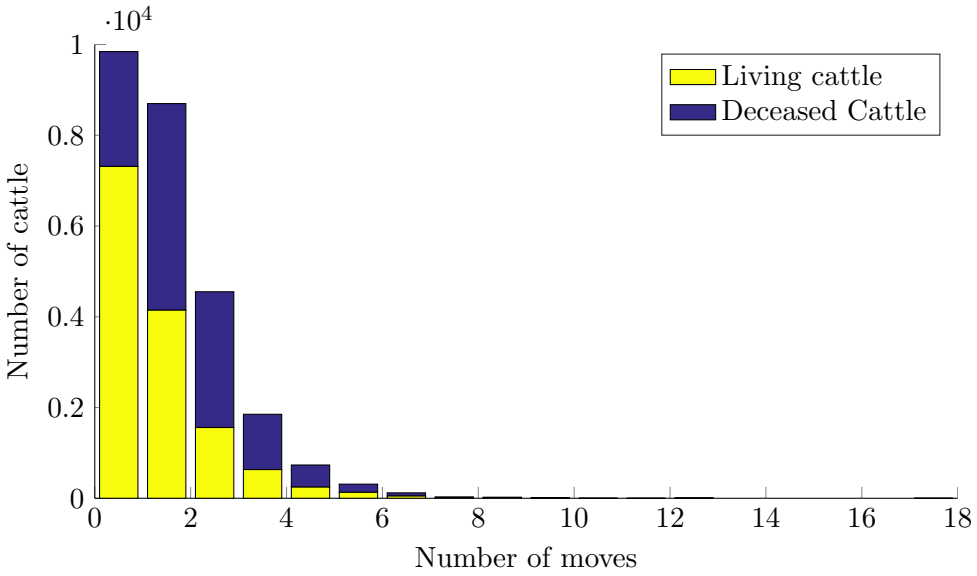

Figure 5.3: Number of movements made by deceased cattle (blue) and those still alive (yellow).

the parameters are bounded between 0 and 1 whereas other parameters such as the recovery and vaccination periods have ranges which are theoretically unbounded.

5.3 Farm level data

As previously mentioned the quality of the movement data is very good after September 2006. For simplicity we take the start of 2007 to begin any disease modelling but if records pre 2007 are found they have been included in a specific cow's history. We describe the procedure for cleaning the data for farm 8,149 below; the same procedures were followed for farm 25,372.

We consider the history (birth, movement and death) of all cattle born, moved or died on farm 8,149. Records exist for 26,171 cattle in this time, of these 12,054 have both a birth and a death, 12,526 have only a birth record and 1,277 have only a death record, and 314 have neither. We refer to births, deaths and movements collectively as events.

The event data was cleaned as described in chapter 2, removing duplications and ensuring moves occur in the correct order both temporally and spatially. Figure 5.3 plots the distribution of the number of movements that both the deceased and living cattle have made. The majority of cattle have only one or two movements throughout their lifetimes.

Figure 5.4 shows the lifespans of the 12,054 cattle which had both a birth

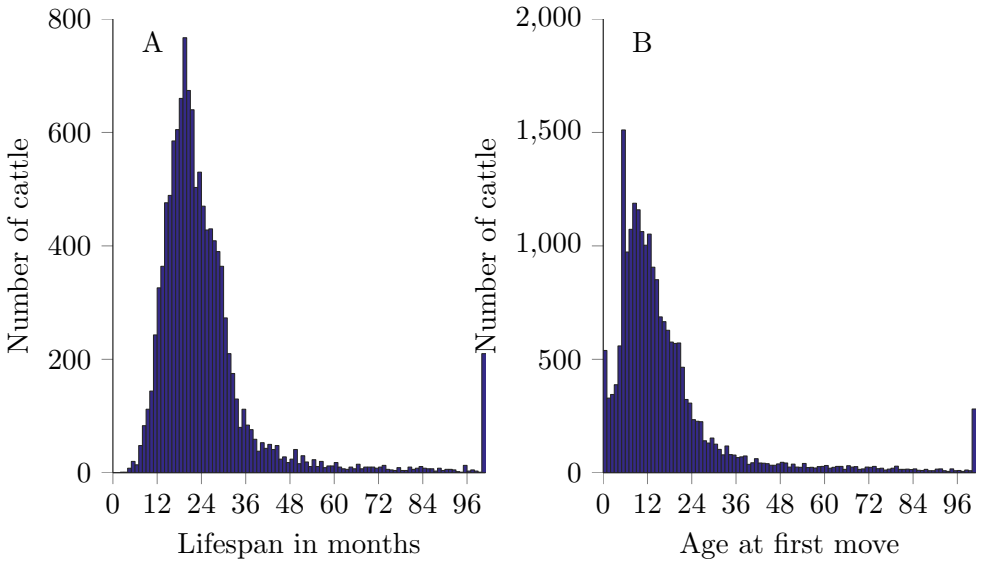

Figure 5.4: The lifespan and age of cattle at first move in months for farm 8,149

and death record. If we express age in months (1 month = 30 days) then the mean age is 25.0 months, the mode is 19 months and the median is 21 months.

5.3.1 Population

As discussed in chapter 2 the number of deaths is not accurately recorded in the data we received. Indeed, running through the movement data for farm 8,149 the populations grows to around 11,000 by the start of 2010. The population of farm 8,149 ranged from 3600-6114 as given from the outbreak data table 5.2. Another list received for large ruminant populations on a village level from 2010 quotes a figure of 6,539[iv].

In order to obtain a better estimate of the true farm population we can attempt to remove cattle that are likely to have died or moved off the farm. To do this we can use some of the techniques discussed in chapter 2 where we derived the expected age distribution of cattle in Turkey.

One option we have is to randomly remove cattle from the farm and to then assign a random date of removal. However we can make a more informed choice by using the available movement data. There are two types of cattle to consider, those that have been born on the farm and have never left, (we refer to these as natives), and those that have moved to the farm and still remain, (we will call these foreigners). Foreigners may also consist of native cattle that have returned to the

[iv]This is *file4* referred to in section 2.4 of chapter 2.

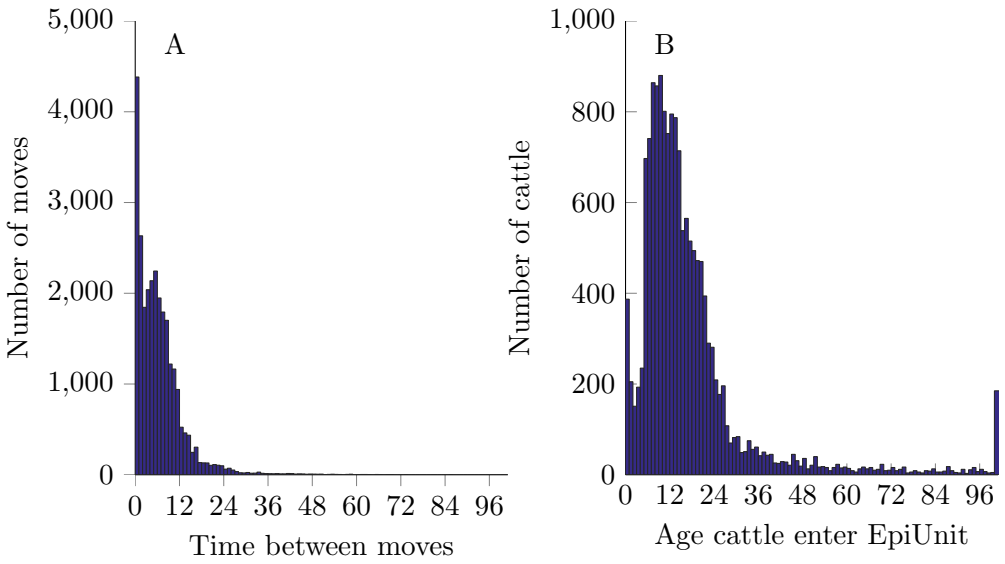

Figure 5.5: The waiting time between moves and the age cattle enter the farm.

farm.

We examined the age distributions at which both native and foreign cattle leave the farm figure 5.6. Preforming a two-sample Kolmogorov-Smirnoff test on these distributions rejects the null hypothesis that these samples are from the same distribution at the 5% confidence level implying the need to treat native and foreign cattle distinctly [v].

To decide which native cattle to remove we examined the age at which cattle born on the farm had left. A normalised cumulative distribution function (CDF) was created using the ages at which native cattle left the farm. This CDF gives the probability that at a given age a native cow has left the farm. The ages of native cattle who have not yet left the farm are then examined. For each of these cattle we assign a probability p that it has in fact left the farm for slaughter based on its age and the previously described CDF. A random number r is drawn; if $p < r$ we keep the cow, if not we remove the cow. In this way we can remove older cattle that have a high probability of having died/left without a record being kept. After deciding to remove cattle we can also allocate dates of removal by again looking at the age of first move for native cattle.

The same process is carried out on foreign cattle in the farm. The only difference is the new removal dates are drawn from the time-between-moves distribution (figure 5.5). Figure 5.7 shows the size of the farm with the original and the

[v]When preforming the two-sample Kolmogrov-Smirnoff test the distributions are created using daily rather than monthly bins, this make the data more continuous.

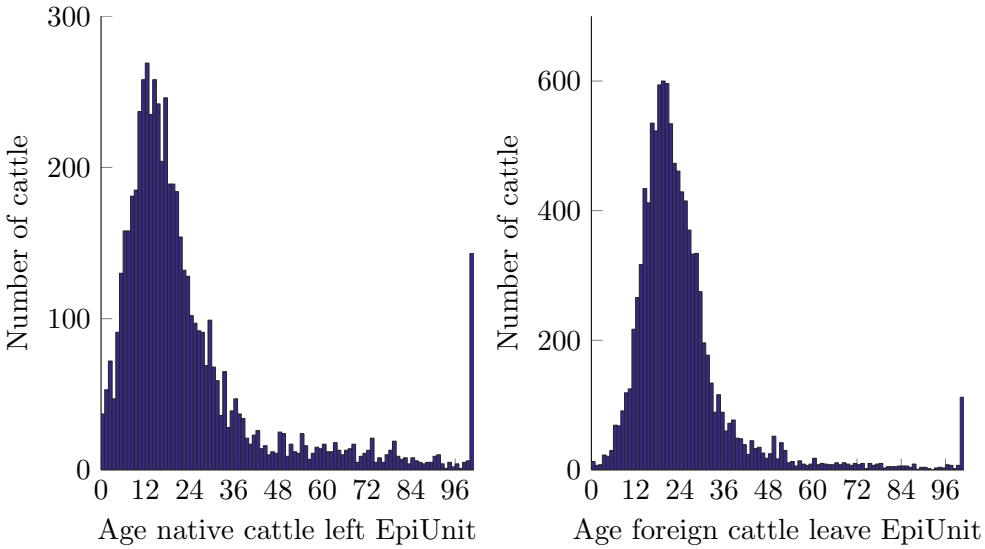

Figure 5.6: The age that cattle born on the farm and cattle that have moved to the farm leave the farm.

smoothed data. Figure 5.8 compares the demography of the original data with the the smoothed data. We can see that the main effect is to remove older cattle from the farm.

5.3.2 Outbreaks

Table 5.2 and table 5.3 are extracts from the outbreak data for farms 8,149 and 25,372 respectively. The province, district and village name are supplied for each record but not the farm number; these are obtained by cross-referencing with the list of farm supplied with the movement data. Each record contains serotype, (if known), as well as: a start, confirmation and end of quarantine date. There are also other categories to contain numbers of cattle, though these are quite often left as zero. The most useful of these are the numbers vaccinated, diseased, suspected and the total population.

The movement data is of highest quality from the end of 2006 onwards and as such we will not try to fit to any outbreaks before 2007. The NSP prevalence parameter will attempt to capture the number of cattle on the farm still in the recovered state from outbreaks prior to 2007.

At this stage of model development we will use a single strain model as the complexities of a multi-strain model will increase the parameter space massively and we do not feel that the data is resolved enough for any model to give meaningful results. We will fit to type-O outbreaks. Type-O has the advantage of being the

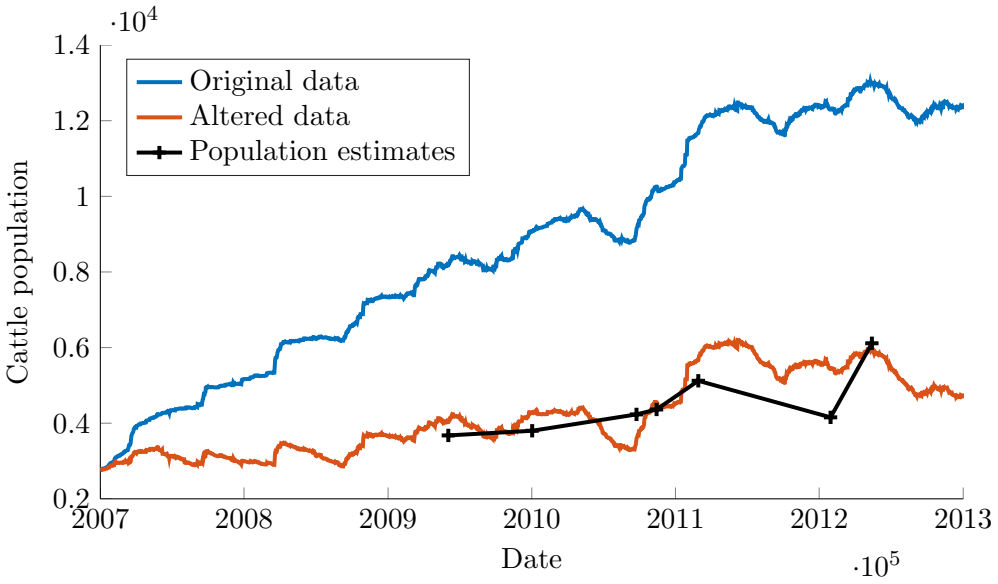

Figure 5.7: The number of cattle on farm 8,149 according to the original data and the altered data. Estimates from the outbreak data of table 5.2 are also shown on the start date of the outbreak.

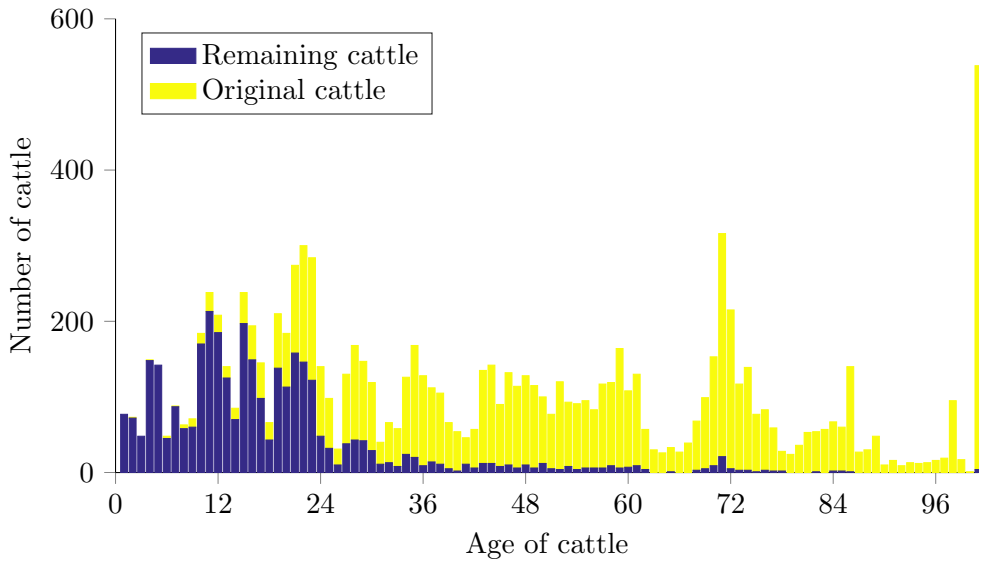

Figure 5.8: Age demographic for cattle on farm 8,149 according to the original data and the altered data.

		Dates			Numbers			
#	type	start	confirmed	q-end	vacc	dis	suspect	pop
1	O	01/05/2002	01/05/2002	10/06/2002	600	47	750	0
2	O	06/07/2005	19/06/2005	19/07/2005	2526	20	350	0
3	O	03/06/2009	16/06/2009	10/07/2009	3318	30	0	3677
4	A	02/01/2010	22/01/2010	26/02/2010	3988	23	247	3800
5	O	24/09/2010	06/10/2010	27/10/2010	4317	120	0	4233
6	O	14/11/2010	26/11/2010	04/01/2011	2659	72	72	4363
7	O	28/02/2011	14/03/2011	28/03/2011	3426	8	292	5122
8	Asia-1	30/01/2012	21/02/2012	03/04/2012	3613	155	0	4155
9	Asia-1	14/05/2012	22/05/2012	18/06/2012	3613	3	0	6114

Table 5.2: Recorded outbreaks for farm 8149. We will use outbreaks #3, #5, #6 and #7 for model parametrisation.

most prevalent strain in the farms considered and that most challenge experiments of FMD to naïve and vaccinated cattle have been performed with type-O [Mardones et al., 2010]. As there is no strong evidence to suggest that having being infected by one strain of FMD infers immunity or increases susceptibility to another it should be possible to parametrise a single-strain model even if there are multiple strains in the population.

For most of the outbreaks considered the vaccinated number is close to the population number quoted in the data with any disparity presumably coming from calves that were considered to young to be vaccinated or were sufficiently isolated from the holdings housing infected cattle. We have no knowledge of which holdings within the farm became infected. During the simulation we will attempt to vaccinate every susceptible cow and give booster vaccination to those in a vaccinated state. In this way recovered cattle and those with maternally derived immunity will not be affected by targeted or routine vaccination campaigns.

We will use the sum of the diseased and suspected columns as the vector of total outbreak sizes $\mathbf{I_D}$ for computation of the summary statistic. For farm 25,372 we ignore outbreak #4 as it has a total size of 1. This outbreak is recorded as still being under quarantine during September 2007 and as such will be subject to routine vaccination so missing the extra targeted vaccination that would have been triggered by an outbreak should not effect the overall results. We also include outbreak #7 into the $\mathbf{I_D}$ vector for farm 25,372. This outbreak was un-typed but including it allows us to fit to three rather than two data points.

		Dates			Numbers			
#	type	start	confirmed	q-end	vacc	dis	suspect	pop
1	?	05/08/2005	01/05/2005	19/09/2005	260	38	285	0
2	A	01/03/2006	20/02/2006	03/05/2006	332	20	189	0
3	A	17/07/2006	18/07/2006	04/09/2006	377	21	30	400
4	O	17/08/2007	24/08/2007	26/09/2007	99	1	0	100
5	O	31/05/2010	31/05/2010	30/06/2010	400	9	20	450
6	O	27/08/2010	01/09/2010	19/10/2010	400	5	9	456
7	?	19/02/2011	22/02/2011	11/04/2011	520	4	0	516
8	A	11/07/2011	11/07/2011	05/09/2011	520	0	0	643

Table 5.3: Recorded outbreaks for farm 25,372. We will use outbreaks #5,#6 and #7 for model parametrisation.

5.3.3 Preparing for simulations

As the outbreak in farm 8149 after 2007 are between June 2009 and May 2012 we run the events without any disease from 2007 to May 2009 to give us the cattle on the farm once the outbreaks began. As the mean age of cattle is slightly over two years having run the events data for over two years should accurately reflect the initial cattle on the farm though there is still the possibility that we have underestimated the number of older cattle.

The outbreak data contains a count of the number of cattle infected with FMD and those suspected of being infected with FMD. We err on the side of caution and sum these figures together. There are four outbreaks of type-O FMD on farm 8149 from 2011 to 2012. It is these outbreaks we will fit to.

5.4 Results

5.4.1 Initial results for farm 8149

Initially we ran the ABC scheme for farm 8149. We used a decreasing set of tolerances starting at 0.95 and reduce it to 0.65. We would have liked to reduce this tolerance even further but our efforts were dictated by simulation time.

We used the metric $\|\mathbf{I}_M^{(k)} - \mathbf{I}\| < \epsilon^{(k)}\mathbf{I}_D$ as our summary statistic accepting parameter values that satisfied this condition. However when we used these parameter values to run simulations we found that epidemics larger and continued for longer in the simulations than the data.

This result is due to our summary statistic being insufficient. It is not enough to enforce that the number infected during the outbreak period was within the

tolerance but that also the outbreak has completely died. The new metric becomes

$$(\|\mathbf{I}_M^{(k)} - \mathbf{I}_D\| < \epsilon^{(k)}\mathbf{I}_D) \wedge \mathbf{i}^k(T_Q) = 0 \tag{5.7}$$

where $\mathbf{i}(t)$ is the proportion of cattle in the infected class at time and T_Q indicates the date the quarantine from the farm was lifted.

5.4.2 Parameter inference for farm 25,372

We repeated the ABC process using farm 25,372. We also altered the range of tolerances to begin at 1.5:1.25:1.00 then reducing in steps of 0.05 to 0.60. We only begin to get realistic values for the parameters when ϵ is less than 1 but increasing the range requires less simulation runs than beginning at a finer tolerance.

The results for the main epidemiological parameters; β, σ^{-1}, γ^{-1},w^{-1} and ν^{-1} are shown in figure 5.9 for $\epsilon = 1.5$ and in figure 5.10 for $\epsilon = 0.60$. In the usual manner for analyses of this kind of parameter inference; the off diagonal subplots of these figures are scatter plots of these parameters with the median point highlighted for each while diagonal subplots show the distributions for these parameters.

For $\epsilon = 1.5$ the latent and infectious periods closely follow the prior distributions. The recovery period and the vaccination period both reach the bounds set in the prior but behave very differently; the recovery period does not show any clear peak while the bulk of the vaccination period lies between 200 and 360 days. The rate of transmission, β, has a narrow range of values with the majority of particles being below 0.5 with the maximum value of any particle being 4.04 well below the prior upper bound of 10.

When the tolerance has been reduced to $\epsilon = 0.60$ the distributions for these parameters have changed. The latent and infectious periods have moved away from the priors with the infectious period moving to a much narrower distribution with a median of γ^{-1}=1.70 days while the latent period has actually flattened out with the bulk lying between 0 and 8 days but with a median of σ^{-1} =4.03 days.

The recovery period now offers a bimodal distribution with a minor peak near w^{-1}=250 days and a higher peak near w^{-1}=550 days. The vaccination period has grown beyond its prior upper-bound but has developed a peak around ν^{-1}=260 days which is below the prior upper-bound of 360.

The transmission rate has moved away from the narrow focus near $\beta = 0.5$ and now has a wider distribution around 1. This overall increase in β, σ^{-1}, and $\nu^{=1}$ correspond to increasing the probability of larger epidemics while γ^{-1} decreasing corresponds to a decreased probability of there being infectious after the outbreak

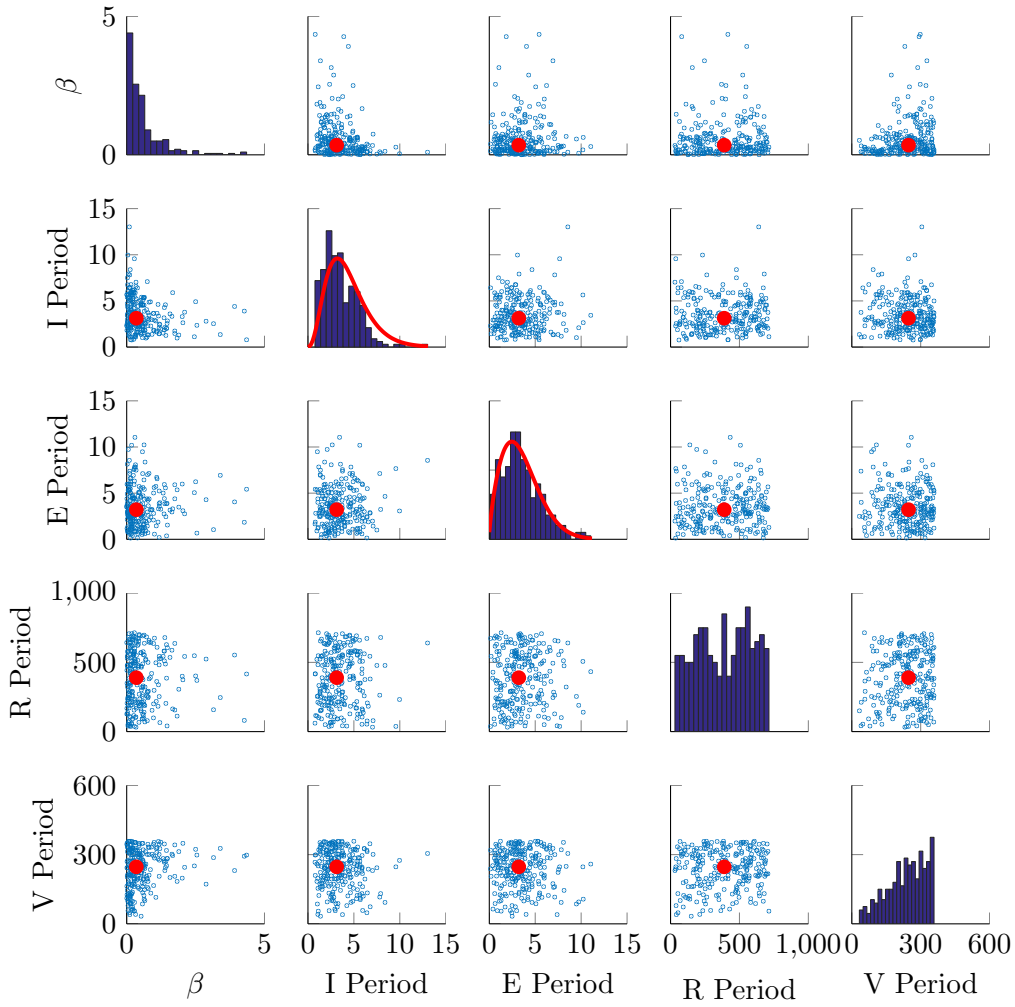

Figure 5.9: ABC results for farm 25,372 using the updated summary statistic from equation 5.7 with tolerance of 1.5. The plots on the diagonal show the distributions for β and the average periods for the E, I, R and V states. The off diagonal elements are the scatter plots for these same parameters, the median values for each pair of parameters is marked as a red dot.

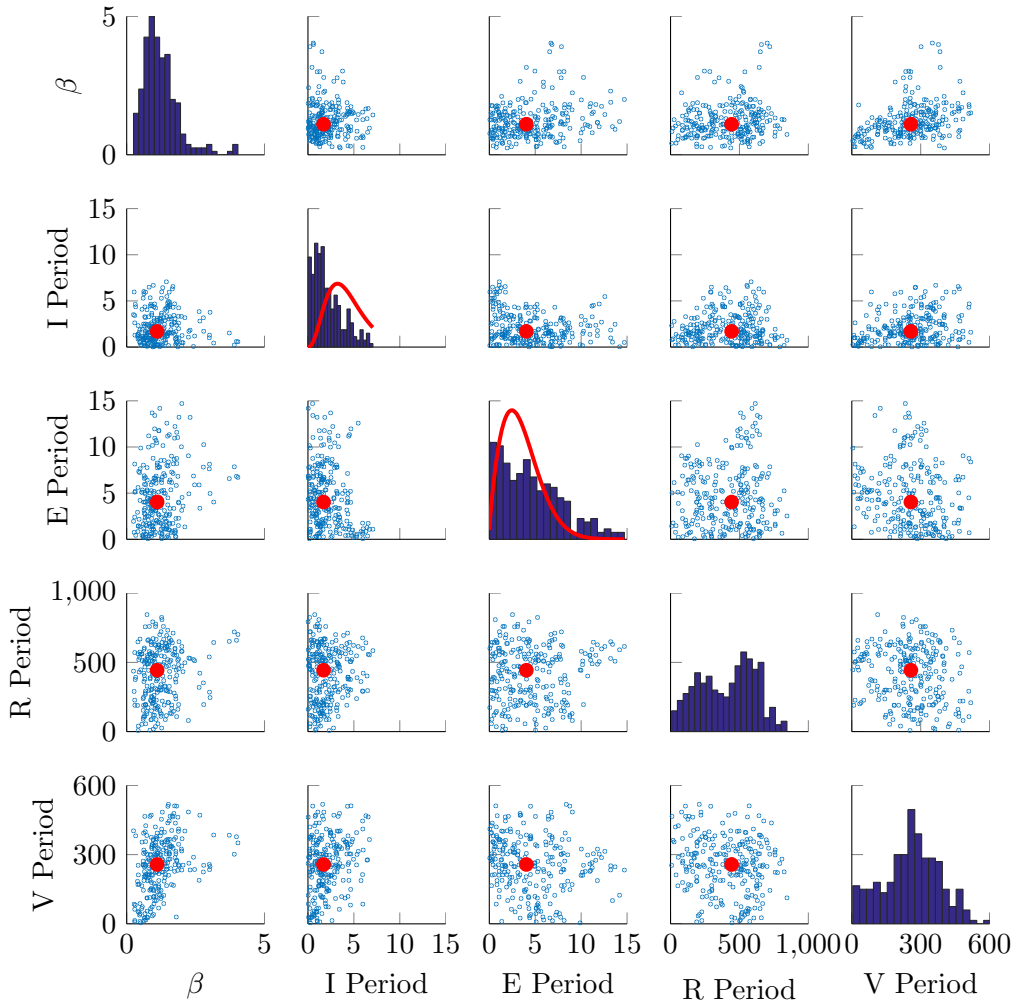

Figure 5.10: ABC results for farm 25,372 using the updated summary statistic from equation 5.7 with tolerance of 0.60. The plots on the diagonal show the distributions for β and the average periods for the E, I, R and V states. The off diagonal elements are the scatter plots for these same parameters, the median values for each pair of parameters is marked as a red dot.

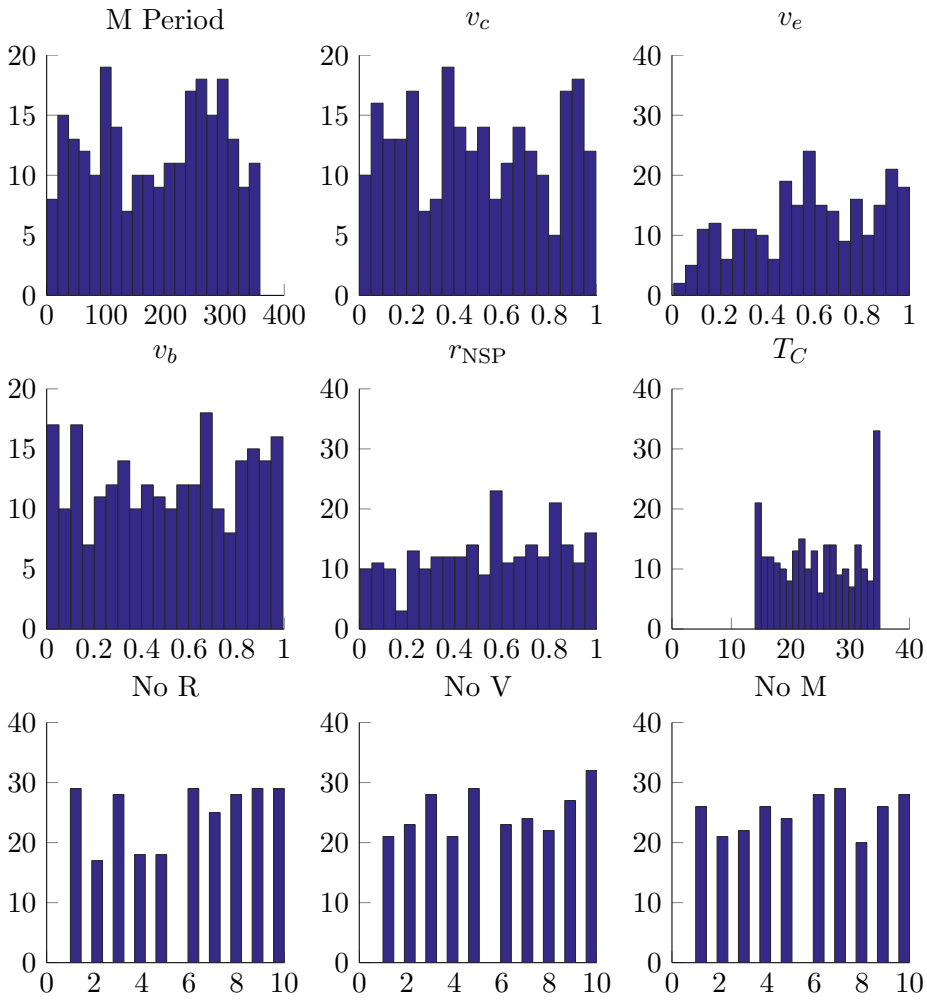

Figure 5.11: Distributions for parameters not described by figure 5.9 with $\epsilon = 1.5$.

should have ended.

For the other parameters we have not drawn scatter plots but instead have plotted there distributions in figure 5.11 for $\epsilon = 1.5$ and in figure 5.12 for $\epsilon = 0.60$.

Of these parameters the two that show the most defined distributions at the strictest tolerance are the period for maternally derived immunity κ^{-1} and the time between the outbreak starting and the reactive vaccination being implemented - time to control, T_C.

The T_C has evolved from a relatively flat distribution which favoured the prior bounds of 14 and 35 days to a distribution bordering on bi-modal with a well defined peak near 20 days comprising the bulk of the particles and a much smaller secondary peak centred around 30 days.

We can also examine how the distributions have changed as the tolerance

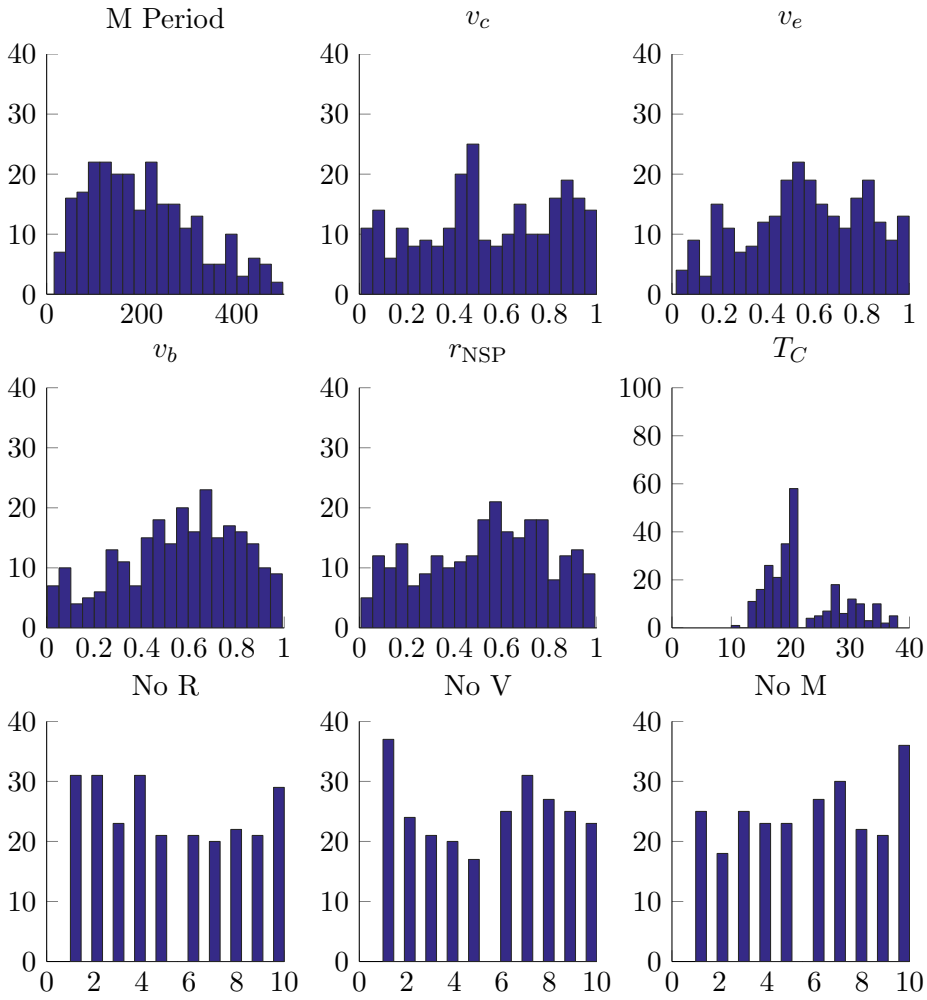

Figure 5.12: Distributions for parameters not described by figure 5.10 with $\epsilon = 0.60$.

has been made stricter by looking at the cumulative distribution for each of the parameters at each tolerance. We plot this for $\epsilon \in \{1.5, 1, 0.8, 0.6\}$ in figure 5.13.

5.4.3 Simulation results using farm 25,372

Now that we have distributions for parameters at a reasonably strict tolerance we can simulate outbreaks sampling from these distributions.

Initially we choose the median value for each parameter and run 10,000 simulations. We limit ourselves to only analysing the simulations that took off for all three reported outbreaks. This limits us to 3,504 of the 10,000 simulations. We also conducted a second run of 10,000 simulations were for each simulation parameters were sampled[vi] from the distributions of figures 5.10 and 5.11. Of these simulations 3,082 had a take-off for each of the three outbreaks.

In our simulations some outbreaks will die out quickly while others will continue on beyond the duration of the outbreak as recorded in the data. For each of the three outbreaks we analyse the simulation results by counting the number of cattle that became infected during the period defined by the outbreak data. This gives us a direct comparison with the data.

The results of this analysis are shown in box-plots given in figure 5.14. In each of these plots we have marked the size of the outbreak as recorded in the data. The results are varied depending on the outbreak. For the first outbreak the simulations underestimate the data with the reported outbreak size lying beyond the 95th percentile of the simulated outbreaks. The second outbreak yields far better results with the recorded outbreak size falling within the interquartile range. The third outbreak also yields a good result though slightly overestimates the size of the outbreak with the recorded outbreak size lying just below the interquartile range.

Similar results are obtained when we sample from the entire parameter distribution. Sampling from the entire distribution increases the variance in final epidemic size for each of the outbreaks. The median values however have not differed noticeably; for the first two outbreaks the median dropped from 6 to 5 and 8 to 7 respectively whilst the median size has remained at 10.

While we unfortunately do not have data pertaining to the way in which individual outbreaks unfolded we can analyse the simulation data in a temporal manner. In figure 5.15 we have plotted the median and mean number of infectious cattle through time. The interquartile range of the number infectious on any given day is also shown as are the duration of the reported outbreaks.

[vi] A sample in this case would be one of the 250 particles found during the ABC-PRC scheme.

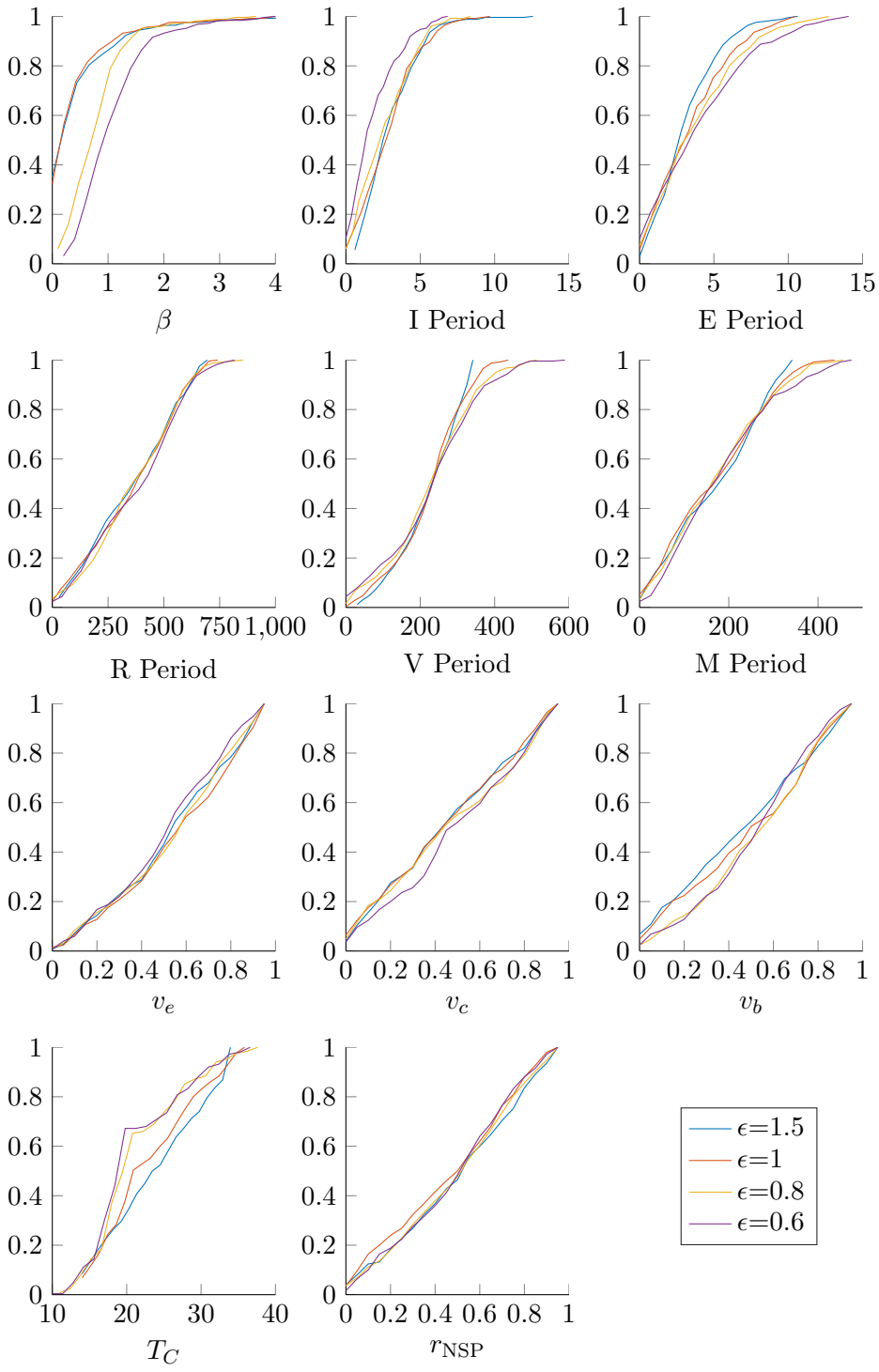

Figure 5.13: Cumulative distributions for various parameters at different tolerances from the ABC scheme implemented on farm 25,372.

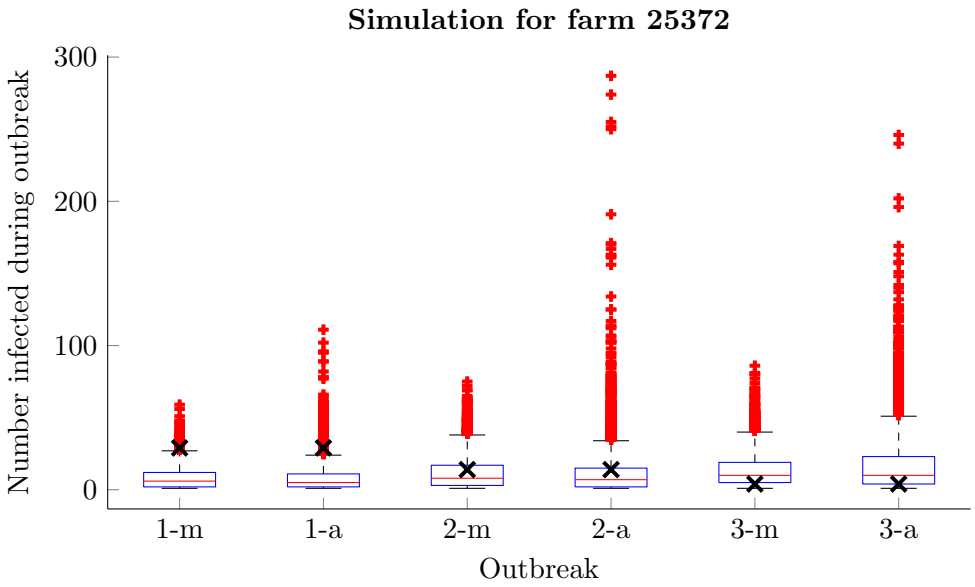

Figure 5.14: Box plots for simulation results for farm 25,372. The outbreak size from the data is given is marked with an x for each of the three outbreaks. We compare the results when only the median parameters are used (m) and when all parameters are sampled from (a). Only simulations in which all four farms had an outbreak take off are included (3,504/10,000 for m and 3,082/10,000 for a).

From this figure we can see that for the first two outbreaks epidemics within the interquartile range successfully take-off and die out during the outbreak duration. The third outbreak tends to persist past the outbreak duration given in the data but is eventually controlled by a combination of the reactive and routine vaccination.

The same analysis is conducted for the simulations carried out when sampling from the full parameter distribution. The higher variance noted in the box plot analysis is seen in the temporal data but the epidemics follow the same pattern of being controlled within the outbreak duration for the first two outbreaks but persist longer for the third outbreak. This figure is given in appendix D figure D.1.

5.4.4 Simulation results for farm 8,149

Due to computational constraints we were unable to parametrise farm 8,149 directly using its outbreak history. However, we can use the parameters found for farm 25,372 to run simulations on farm 8,149. A priori we do not know what kind of results this may yield. Ideally the parameters from one farm would apply to every other farm; it is more likely that the size and production differences between the two farms mean

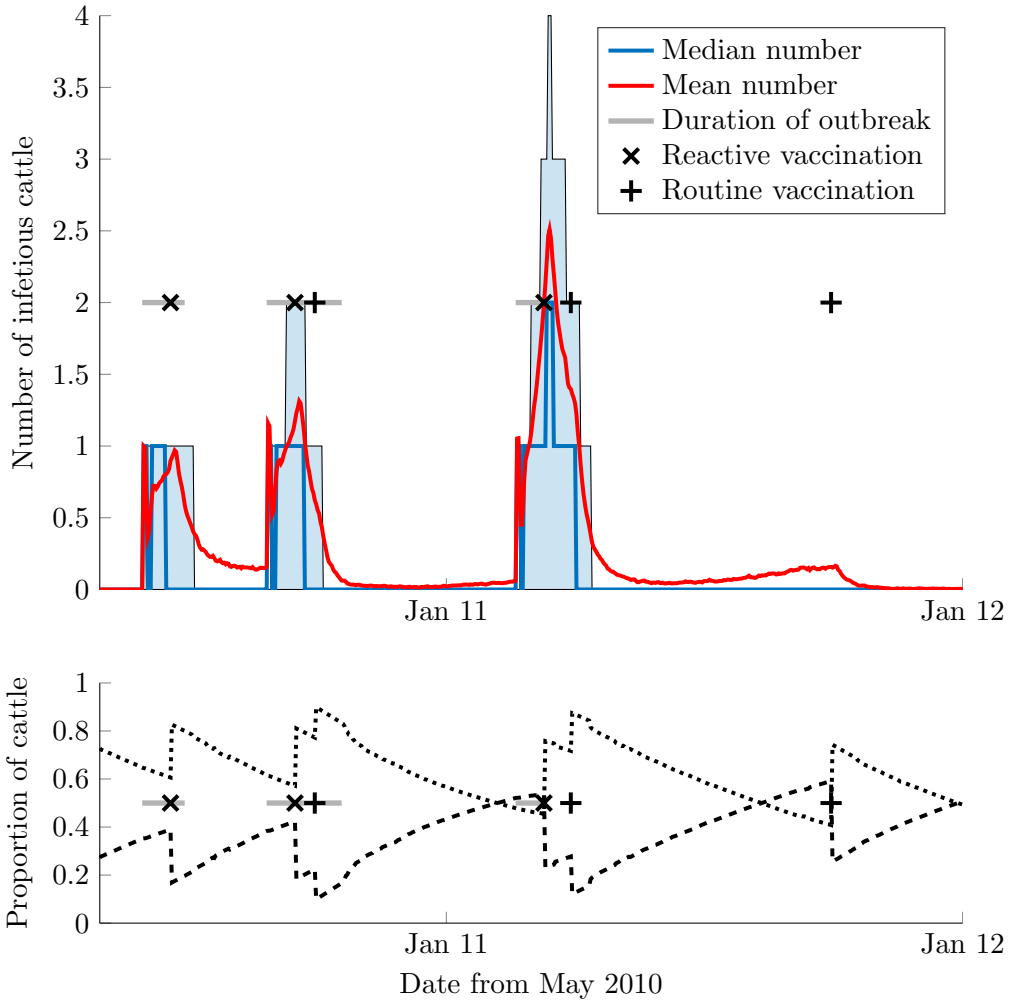

Figure 5.15: Simulation results for farm 25,372 using the median parameter values found using ABC with tolerance $\epsilon = 0.60$. The mean and median number of infectious cattle are plotted each day in the upper figure with the interquartile range shaded in grey. The duration of the outbreaks are also shown along with the dates of reactive and routine vaccination. The lower figure shows the proportion of immune (dotted line) and susceptible (dashed line). Only simulations were all three farms had an outbreak take off are included (3,504/10,000)

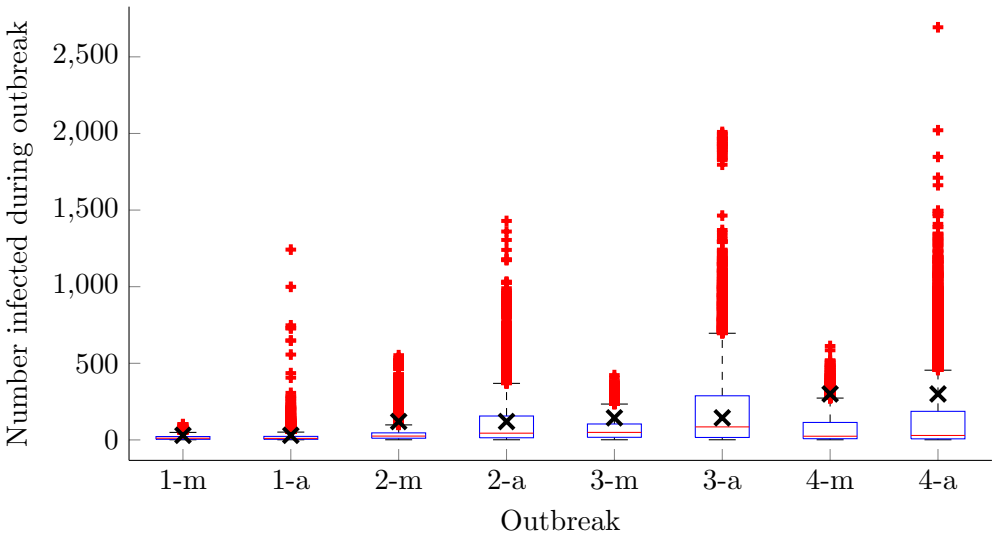

Figure 5.16: Box plots for simulation results for farm 8,149 using parameter values found by using ABC-PRC in farm 25,372. The outbreak size from the data is given is marked with an x for each of the four outbreaks. We compare the results when only the median parameters are used (m) and when all parameters are sampled from (a).Only simulations in which all four farms had an outbreak take off are included (4,540/10,000 for m and 4,116/10,000 for a).

that parameter distributions are best suited to different classes of farm and each would lend itself to a different set of parameter distributions.

We once again run 10,000 simulations using median parameter values and a further 10,000 sampling from the full distribution. We are once again looking to see if the simulations can capture the size and duration of the data.

We make a slight alteration to the routine vaccination. According to our schedule a vaccination round should take place on the 30th September 2010, however the second outbreak begins only six days earlier on the 24th September. As shown in appendix D (figure D.2) this routine vaccination stifles the outbreak preventing it from taking off. As have no precise information on the routine vaccination dates we have omitted the vaccination on the 30th September to allow the outbreak to develop more naturally.

We find that only considering the median parameter values underestimates the size of all four outbreaks (figure 5.16). Sampling from the full parameter distribution improves the estimates for epidemic size in each of the four outbreaks. Most notably we see that simulation estimates encompass the outbreak size from the data in the interquartile range for second and third outbreaks.

The temporal analysis of these simulations is plotted in figure 5.17 (the results

with the median parameters are given in figure D.3).; here we see that the ability for the model to capture the data is very much outbreak dependent. We see that within the interquartile range the first outbreak is controlled within the outbreak duration. However, on average the number of infectious cattle after the outbreak duration is greater than one causing secondary and tertiary peaks in the number of infectious cattle which are controlled by the reactive vaccination. By the time of the second outbreak the tertiary peak has diminished and on average there is less than one infected cattle preceding the second outbreak.

The dates of the second and third outbreak are close together suggesting the possibility that the disease was not fully eradicated from the second outbreak. This is reflected in the simulations were we see that the median number of infectious cattle does not drop to zero between the second and third outbreaks. In future simulation work it may therefore be worth while investigating the effect of combining these outbreaks.

5.4.5 Model consistency

While time did not permit us to fully explore the parametrisation of farm 8,149 we can check to see if the model is consistent within itself. To do this we select a simulation from those simulated on farm 25,372 and refit the model to the simulated data. Thus we will know in advance the parameter values used to generate the outbreak data and can judge if the posterior distributions give reasonable estimates of these parameters. We can also compare the new distributions with those we previously found.

We chose a simulation generated using the median parameter values found for farm 25,372 in section 5.4.2. This simulation had outbreaks of sizes $\{11, 19, 2\}$, the original outbreak data had sizes of $\{29, 14, 4\}$. As the simulated data has outbreaks that are smaller in number we would expect our new posterior distributions to underestimate some parameters such as β and infectious period, and vaccine efficacy.

We run the inference in the same way as before using the original prior distributions and starting with a tolerance of $\epsilon = 1.5$ which we lower to $\epsilon = 0.60$. The posterior distributions for the model fitted to the simulated data are given in figure 5.18, in which we indicate the means of these distributions and the parameters used to simulate the data.

We can compare these distributions with those of figure 5.10 and figure 5.12 by first checking if they are statistically the same as indicated by the two sample Kolmogorov-Smirnoff test[vii]. Performing this test between every pair of distributions

[vii] We construct the cumulative distribution functions using the raw parameter values not their

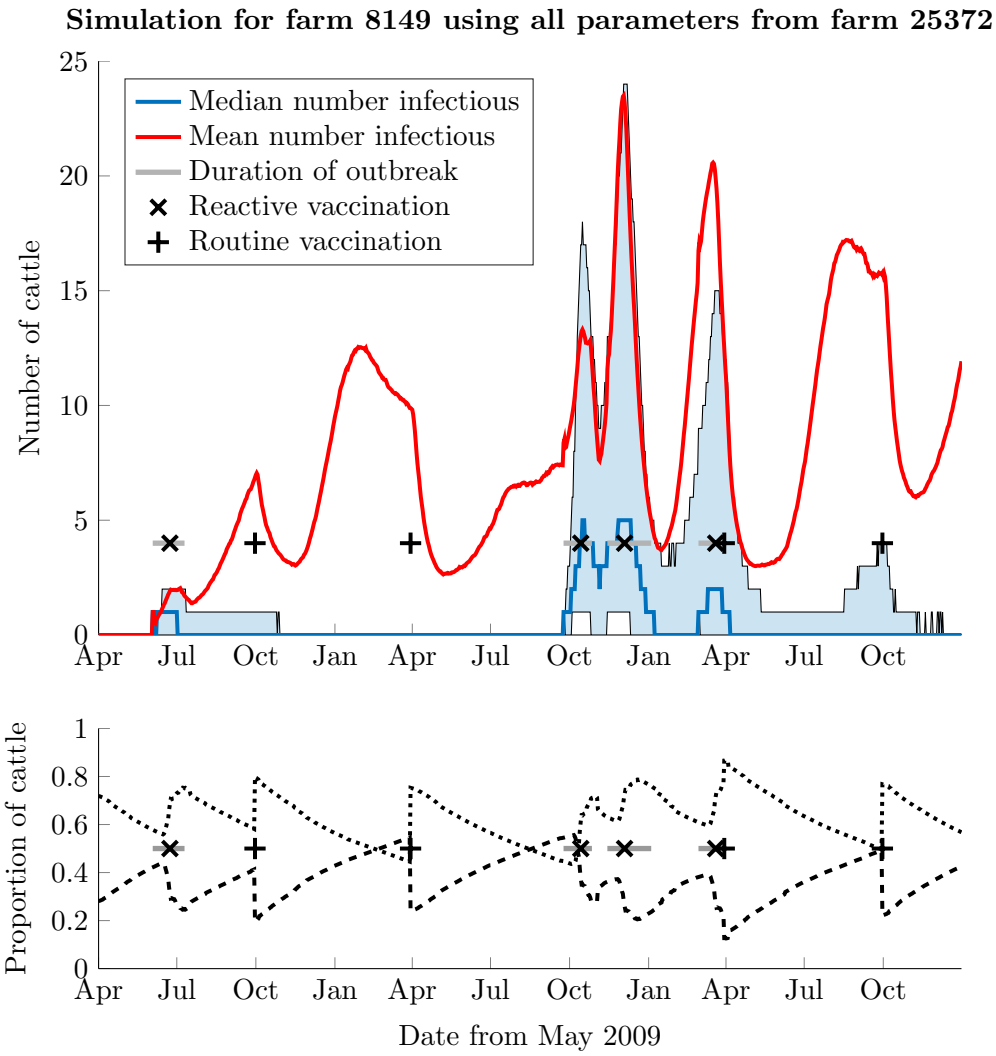

Figure 5.17: Simulation results for farm 8,149 using the parameters drawn from the distributions values found using ABC with tolerance $\epsilon = 0.60$ for farm 25,372. The mean and median number of infectious cattle are plotted each day in the upper figure. The duration of the outbreaks are also shown along with the dates of reactive and routine vaccination. The lower figure shows the proportion of immune (dotted line) and susceptible (dashed line). Only simulations were all three farms had an outbreak take off are included (5004/10000)

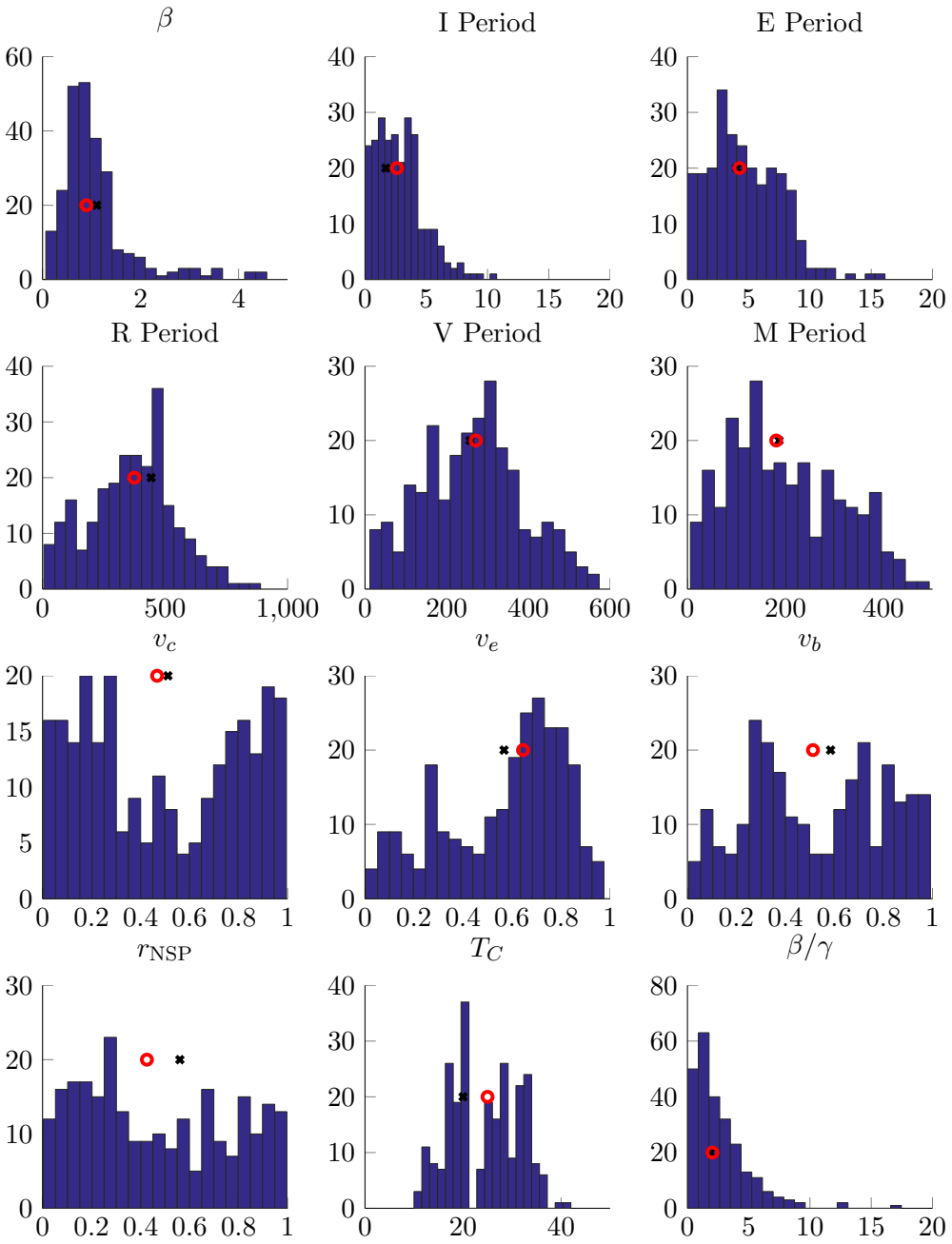

Figure 5.18: Posterior distributions at $\epsilon = 0.60$ for the model fitted on simulated data. The parameters used to simulate the data are indicated by black crosses while the red circles indicate the median parameter values of the distributions.

we find that the distributions for the exposed period, the vaccination period, and the period of maternally derived immunity pass the test at the 5% confidence level as do the distributions for the number of states in the R, V, and M classes. The test is failed by the other parameter distributions rejecting the null hypothesis that they come from the same distribution.

For those parameter values that passed the Kolmogorov-Smirnoff test we can see from figure 5.18 that the median parameter values are very similar. The distributions for the β and the infectious period both failed the Kolmogorov-Smirnoff test however, if we multiply these values together to form a rough estimate of R_0[viii] we find that this new distribution does pass the Kolmogorov-Smirnoff test and that median parameter values lie very close to one another. This would indicate that for the main epidemiological parameters the model is consistent.

We will now discuss the remaining parameters that failed the Kolmogorov-Smirnoff test. The recovery period was underestimated in the refitted inference and failed to reproduce the bimodal distribution previously attained. However, the median value was over a year and as shown in figure 5.3 B over 95% of cattle move off the farm within 12 months as such this parameter is likely to have little impact on disease dynamics.

There are three parameters that determine the efficacy of vaccination: the coverage v_c which effects cattle imported into the farm, the effectiveness v_e which determines whether or not a cow on the farm will become vaccinated, and the boost effectiveness v_b which moves vaccinated cattle back into the first vaccinated state. However, while intuitively we would expect these parameters to be correlated, in practice we could not find such a correlation and revert to a more qualitative appraisal of their respective distributions. None of these distributions had clearly defined peaks as such their role in the model is uncertain. Clearly vaccination is very important to the model and as demonstrated in the simulation is necessary to control the outbreak. What is less clear is how much vaccination there needs to be, perhaps vaccination efficacy only needs a minimum value of 0.6, say to control the outbreak after which any increases play no part and variation beyond this threshold is merely noise. Before further development of this model it would be wise to tease out the effect of these parameters by running simulations where other parameters are kept constant. This may allow us to remove parameters that do not have an effect.

binned equivalents.

[viii] A through calculation of R_0 would include the exposed period, take into account mortality and assume a closed population. We do not have a closed population as cattle are moving on and off the farm at a time-scale comparable with the outbreak durations.

5.5 Discussion and future directions

The decisions we made in building our model have come from the data we have available. We have comprehensive data on cattle movement data allowing us to know when individual cattle moved on and off the farm. This is important as work by Keeling et al. [2010] found that loss of individual identity in models of FMD in the UK caused larger outbreaks compared to models where cattle identity was maintained.

The size of the farm is also directly incorporated into our model in the transmission probability. Our methodology to apply probable dates of death to cattle has allowed estimates of the farm size that align well with the estimates from the various data sources available to us. As such we can be confident in our estimates of the size of the farm.

Aspects of the model could be improved with the availability of certain amounts of data. Most obviously would be vaccination data. A greater knowledge of when routine and reactive vaccination took place and which cattle where vaccinated would reduce the parameter space of the model and increase confidence in the simulation results. If vaccination data could be coupled with knowledge of exactly which cattle became infected during an outbreak we could better determine the efficacy of the vaccine used at the time. Such data would also increase our ability to estimate waning immunity and maternally derived immunity.

While the simulations carried out using the median values from the parameter distribution in farm 8,149 underestimate the size of the outbreaks it is encouraging that sampling from the full distribution greatly improves these estimates. This suggests that the parameters distribution for farm 25,372 can be applied to other farms.

We could extend this work by separately parametrising farm 8,149 using the same prior distributions as for farm 25,372. We could also use the posterior distributions we obtained for farm 25,372 as prior distributions for the inference scheme applied to 8,149. If the distributions found from these two methods are statistically the same then this would be a powerful result; if it is notably quicker to parametrise a large farm with priors obtained as posteriors from a small farm than it is to parametrise the large farm from scratch.

The available data dictates the construction of the model. A well constructed model should be able to reproduce the observed data using the minimum number of parameters. Aspects of our model still need to be tested to ensure that the results we have found are not artefacts of the model construction. For instance we have

not varied the start date of any of the epidemics nor have we experimented with the dates of the routine vaccination.

We have also only applied our model to farms suffering from multiple outbreaks of type-O FMD. This allowed us to use priors derived from the literature which tends to predominately focus on type-O. It would be informative to extend the model to farms that have been inflicted by types-A and Asia-1 also. However, we could find no farms that were inflicted with sequential cases of one of these strains without a case of the other strain of type-0 in between. As it stands the model is not flexible enough to allow for multiple strains of FMD unless we make the bold (and almost certainly mistaken) assumption that immunity from one strain infers immunity from all strains.

Despite the limitations discussed above we have successfully developed and parametrised a model that captures key epidemiological components of a succession of outbreaks in two farms. To our knowledge this is the first study that has been able to do this and puts us in good stead to further improve and refine the in-farm model before incorporating it into a more ambitious national scale model.

Chapter 6

Conclusions and horizons

> "The road and the tale have both been long, would you not say so? The trip has been long and the cost has been high... but no great thing was ever easily attained. A long tale, like a tall Tower, must be built a stone at a time."
>
> Stephen King, The Dark Tower [2004]

This thesis has had a broad scope. The focus has been on the modelling of FMD and the analysis of livestock movement networks. Many aspects of these two themes remain unexplored. In this chapter we will summarise what we have done and how this work can be built upon, laying out the foundations for a full national scale model of FMD in Turkey.

6.1 Data, network analysis and community detection

The Turkish Veterinary authorities provided us with a wealth of data from their administrative systems. The processing and cleaning of this data was a huge but worthwhile task. In cleaning the data we learned about its strengths and weaknesses.

The movement data was very comprehensive and while there were errors we were able to impose rules that fixed the vast majority of these errors and ensured that every cattle in the database had a sequential history of movements. The distributions of the number of movements per cow and the waiting times between movements seemed sensible and reassured us that through cleaning the data we had not introduced our own biases.

A clear failing of the data we received was the under-reporting of deaths. However, through analysis of the cattle which had both birth and death records we were able to construct distributions for life expectancy and thus provide plausible

dates of death for the remaining cattle. This process was validated when we applied it to both the farms we analysed in chapter 5 where our estimates for the sizes of the farms matched reasonably well with the estimates provided to us through various data sources.

A limitation of the data from both an analysis and a model construction view is the lack of information about holdings. Each farm contains a multitude of holdings with the distribution ranging from one holding to over 300 in some cases. While we are able to estimate the number of cattle on a farm it is challenging to estimate the number of individual holdings. While we have positional data for the farm we have no idea of the area the farm covers, with no information at all about where holdings sit within a farm. We know that farming practices differ around the country with intensive fattening and dairy farms in the West compared to communal grazing in the East. As such the make-up of the holdings and their area within a farm is important as it gives us information about factors which could influence the spread of disease.

If a farm was comprised of high density holdings that were separated we may reasonably expect any disease incursions to spread rapidly within the holding but that spread outside of the holding within the farm may be more easily controlled. If on the other hand, the holdings were actually small households with communal grazing than any incursion of disease would have a higher probability of spreading throughout the farm.

While the make-up of individual farms is clearly important for understanding within-herd dynamics of FMD, aggregating movements at the farm level meant that this limitation had no effect on our network analysis. Indeed even with full knowledge of holding location we would have aggregated the movements anyway as applying most measures (in particular shortest-path and modularity optimisation) to a network of over two million nodes is infeasible. The farm level network of 55,000 nodes is far more manageable and statistics measured at this level are more likely to give meaningful results with less noise caused by the multitude of small holdings which may lie dormant in the database for months and sometimes years.

While we only examined network statics aggregated spatially at the farm level we did consider different temporal resolutions. Bearing in mind that the infectious period of FMD on a farm[i] is generally accepted to be around three to four weeks aggregating movements to the monthly level seemed a natural choice.

Aggregating at the monthly level also allows us to capture seasonal trends that are evident in the movement data and in most of the network statistics. Clear

[i]Not within cattle

increases in movement numbers in the spring and around the time of the Kurban festival were observed in all years that we examined. The size of the giant strongly connected component increased at these times of year, as did the average degree and the betweenness centrality. These measures are indicators for the connectivity of the network. As connectivity is increased disease spread through the network becomes more likely.

Beyond the seasonal trends we also observed a complete shift in many statistics between the periods 2007-2009 and 2010-2012. We observed that the network became more assortative, the proportion of bi-directional links decreased, local clustering decreased and that betweenness centrality increased. All of these change can facilitate disease spread across the network. We also observed the community structure changed between these periods. The reasons for this apparent structural change in the network cannot be explained by the data alone and is probably linked to some change in policy. Going forward we will seek greater guidance from our Turkish colleagues in determining the causes behind our observations.

6.1.1 Partial data

The work of chapter 3 showed that it is indeed possible to make accurate epidemic predictions without full knowledge of the underlying network. We demonstrated that this was heavily dependent on how the network was sampled, with targeting of high degree nodes providing the strongest results. As previously discussed this leaves us in some what of a quandary, we have to sample nodes based on their degree but presumably we do not know their degree before we have sampled them. Our solution to this problem would be to mandate farmers to record the number of trades they have made in one year and use this statistic to choose which farms are fully sampled the following year. This relies on the trade volume of farms staying reasonably similar from year to year.

However if cattle movement records had never been recorded then choosing how to sample farms would be difficult. A suggestion would be to build on the idea of Gilbert et al. [2005] in using meat-demand discrepancy between districts as a proxy for inter-district movement. Once the districts with the highest trading volumes were established then farms within those districts could be sampled based on herd size.

While we found good results for the British cattle network as of yet we have not extended this work to the Turkish cattle. It would be interesting to see how well the results hold and if the point at which epidemic simulations on a sampled network can be considered accurate varies. The structure of the Turkish cattle industry with

farms (epidemiological units) being comprised of multiple holdings may mean that drawing on summary statistics from one year will not lead to accurate sampling in the following year.

6.1.2 Community detection

In performing temporal community detection we sought to bring novel analysis to the cattle movement networks of the UK and Turkey. Our analyses mainly focused on use of the Newman-Girvan null model. In many ways this is a natural null model to use as it is comparable to comparing our community structure to the average structure one would expect to find if a network of the same degree sequence was constructed using the configuration model. Indeed we found some interesting results, highlighting the robust isolation of Scotland in the UK cattle network and observing the strong connection between Ankara and Ezurum in the Turkish network.

However, farming networks are firmly embedded in space and as such we believe more work should be invested in alternative null models that specifically incorporate the distances between nodes. In our implementation of the gravity model proposed by Expert et al. [2011] we found a community structure in the UK with clear difference from that found using the Newman-Girvan null model. Scotland and the North of England were grouped together and a discontiguous community between the South-West and East Anglia was observed. This county level spatial partition also performed well as a predictor for epidemic spread.

Despite these results we limited ourselves to a county level rather than a farm level approach. This was due to computational constraints. For the gravity model the distance between every pair of farms must be recorded which with over 70,000 farms in the UK was too computationally expensive compared to the network approach which typically had a sparse adjacency matrix. However, we are confident that perseverance in the development of new algorithms and methods to overcome these limitations would be worthwhile as we believe that to truly understand the drivers that cause edges to be made in the cattle network requires a deeper understanding of the space in which the network is embedded.

Community detection in networks is relatively new as an analytical technique. It is potentially a powerful tool that can in theory unearth relationships between nodes that standard techniques which rely on producing a statistic such as betweenness or degree cannot. However, in a suitably sparse network the algorithm will always find some community structure, researchers must be careful not to assign too much relevance to these communities without bringing in a more subjective view from experts in the field.

With more and more data being collected and becoming available to researchers new methods for data analysis and understanding this data are being developed. In the future it should be possible for a researcher to layer livestock trade data for different species with, say a transportation network to build a multiplex network. The study of multiplex networks is still in its infancy but as we have shown by performing temporal community detection it is possible to gain insights from these networks on real-world data sets.

6.2 Outbreak data

As mentioned in chapter 2 Gilbert et al. [2005] carried out an analysis of provincial level outbreak data from 1990 to 2002 for Turkey. We would hope to extend their work to the analysis of outbreak data from 2007 to 2012. While Gilbert et al. [2005] used meat-demand discrepancy as a proxy for movement we obviously have this data which could lead to more accurate results. We could also perform the analysis at a district rather than provincial level.

We would hope to further develop a logistic regression model in which we could include the community assignment of a district from a given year (or month depending on chosen temporal resolution) as covariate in the model. This may serve as a higher level variable than the number of movements between districts and offer new insights into the spatial-temporal spread of FMD in Turkey.

There are some issues with inferring too much from the outbreak data. The concept of an outbreak is loosely defined and we have no idea of the scale of under reporting of outbreaks. The Turkish veterinary services carry out a 'sero-survey' of the countries cattle farms in which they test for FMD antibodies to determine the underlying prevalence of the FMD virus in the cattle population. These surveys show high prevalence of the virus in the South-East of the country however very few outbreaks are recorded from this region. This may be to an unwillingness for farms in the region to cooperate with the veterinary services for fear of movement bans or increased vaccination regimes, or it may be that farms in the South-East are less able to identify the symptoms of FMD. Either way attempts to model the disease will be hampered by under reporting. A solution to this would to be to focus our efforts on the West of the country where the disease prevalence is less than the East but outbreaks but more outbreaks are reported indicating a much higher reporting rate.

6.3 A national scale model

In section 5.5 we discussed how to improve and extend the within farm model. We may find in this work that it is most appropriate for different farms to have different parameter values. Spread within a farm may be less likely within a more commercial farm where the holdings are very much separate entities than in a farm which is a essentially a small village with communal grazing. As such the transmission rate β, may vary depending on farm type. Further data about the farms such as production type and if they have other livestock may be necessary in order to categorise the farms. Results of the logistic regression model described above may help with this categorisation.

Once we are confident of the farm level parameters the extension of the in farm model to the national scale is straight forward, if spread is assumed to come only from the movement of cattle. Each cow will belong to a farm and the cow will change disease state as defined by the within farm model of chapter 5. A farm can be seeded with an outbreak and if an infected cow moves to another farm this will act as a seed for that farm. This is a simple model, the only complication arising from the computational time associated with the processing of the movement data and the disease state of 14 million cattle. However, a diligent programmer should be able to overcome these difficulties.

The caveat to this being straight forward arises from the spread of FMD not coming from cattle movements alone but also from local spread which is generally modelled using a spatial kernel [Keeling et al., 2001] as discussed in chapter 1. Simulating the spatial spread by partitioning the country into grids is discussed in [Keeling and Rohani, 2008] and work has been done by Brand et al. [2015] to improve upon computational time of this approach.

Building the spatial kernel into the model should not pose too many difficulties, we would introduce a probability of each infected farm infecting every other farm (the gridding method makes this less computationally exhaustive than it sounds). Once a farm has been infected by spatial spread we would then have to decide how many cattle on that farm would become infected.

It is the parametrisation of the spatial part of the model where we may run into difficulties. The Keeling model [Keeling et al., 2001] modelled the 2001 outbreak post-movement ban, as such no movements were required and the only mechanism for infection was through spatial spread. We will have two mechanisms for spread which will lead to some conflicts.

We could attempt to parametrise the spatial spread by fixing the within

farm parameters and running a Bayesian inference scheme using the 2001 outbreak parameters as prior distributions [Deardon et al., 2010]. However, as no movement ban is in place in Turkey we will have to allow movements between farms which will lead to farms becoming infected that were not actually infected. In this way fitting the spatial spread to the outbreak data directly is flawed as we cannot say which outbreaks came from infected movements or which came from spatial spread.

Disentangling the two methods of transmission may not prove possible. In this case we would have to satisfy ourselves with parameter estimates from previous work and test the model to see what appears realistic. Once this is done then a suite of control measures can be tested, from vaccination with varying efficacy, to movement controls by which ever geographies seem appropriate.

Further complications will arrive when livestock markets are considered. In the UK trade through a market must be reported however no such mandate exists in Turkey. Through conversations with Turkish vets and farms our understanding is that market trading is more common in the East of the country and that cattle can be traded freely without health checks and bio-security measures. The situation in the West is once again different, markets have improved bio-security measures and some markets have on-site slaughter facilities particularly around the Kurban time. Some farmers in the West also voiced their concern with markets and choose to avoid them completely and trade directly with other farmers.

As we demonstrated in chapter 3 inclusion of livestock markets in the UK network increases epidemic size dramatically. Modelling the disease at a national level without any knowledge of markets would be a flawed approach. At the very least the location and throughput of the markets would be known so as to establish some sense of what percentage of trades involved markets in different regions.

The ultimate goal of this work was to build a national scale FMD model which could be used to inform policy and help in the control and eventual eradication of FMD from Turkey while at the same time developing novel methods to analyse livestock movement networks. We have made important first steps towards this goal which we, with the help of other researchers, veterinarians, and policy makers will strive to achieve.

Bibliography

Albert, R. and Barabási, A.-L. 2002. Statistical mechanics of complex networks. *Reviews of Modern Physics* 74:47–97.

Anderson, I. 2002. Foot and Mouth Disease 2001: Lessons to be Learned Inquiry Report, HC888. pp. 1–87.

Anderson, R. M. and May, R. M. 1979a. Population biology of infectious disease: Part I. *Nature* 280:361–367.

Anderson, R. M. and May, R. M. 1979b. Population biology of infectious diseases: Part II. *Nature* 280:455–461.

Anderson, R. M. and May, R. M. 1991. Infectious diseases of humans: Dynamics and control. Oxford University Press, Oxford.

Arenas, A., Danon, L., Díaz-Guilera, A., Gleiser, P. M., and Guimerà, R. 2004. Community analysis in social networks. *European Physical Journal B* 38:373–380.

Arzt, J., Baxt, B., Grubman, M. J., Jackson, T., Juleff, N., Rhyan, J., Rieder, E., Waters, R., and Rodriguez, L. L. 2011a. The Pathogenesis of Foot-and-Mouth Disease II: Viral Pathways in Swine, Small Ruminants, and Wildlife; Myotropism, Chronic Syndromes, and Molecular Virus-Host Interactions. *Transboundary and Emerging Diseases* 58:305–326.

Arzt, J., Juleff, N., Zhang, Z., and Rodriguez, L. L. 2011b. The Pathogenesis of Foot-and-Mouth Disease I: Viral Pathways in Cattle. *Transboundary and Emerging Diseases* 58:291–304.

Bajardi, P., Barrat, A., Savini, L., and Colizza, V. 2012. Optimizing surveillance for livestock disease spreading through animal movements. *Journal of the Royal Society, Interface / the Royal Society* 9:2814–25.

Barabási, A.-L. and Albert, R. 1999. Emergence of Scaling in Random Networks. *Science* 286:509–513.

Barthélemy, M. 2011. Spatial networks. *Physics Reports* 499:1–101.

Bates, T. W., Thurmond, M. C., and Carpenter, T. E. 2003. Description of an epidemic simulation model for use in evaluating strategies to control an outbreak of foot-and-mouth disease. *American Journal of Veterinary Research* 64:195–204.

Bhattacharya, K., Mukherjee, G., and Manna, S. S. 2007. The International Trade Network: weighted network analysis and modelling. *Journal of Statistical Mechanics: Theory and Experiment* 02002:9.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. 2008. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* 2008:P10008.

Boklund, A., Halasa, T., Christiansen, L. E., and Enøe, C. 2013. Comparing control strategies against foot-and-mouth disease: Will vaccination be cost-effective in Denmark? *Preventive Veterinary Medicine* 111:206–219.

Brand, S. P. C., Tildesley, M. J., and Keeling, M. J. 2015. Rapid simulation of spatial epidemics: A spectral method. *Journal of theoretical biology* 370:121–34.

Brooks-Pollock, E. and Keeling, M. J. 2009. Herd size and bovine tuberculosis persistence in cattle farms in Great Britain. *Preventive veterinary medicine* 92:360–5.

Brooks-Pollock, E., Roberts, G. O., and Keeling, M. J. 2014. A dynamic model of bovine tuberculosis spread and control in Great Britain. *Nature* 511:228–31.

Buhnerkempe, M. G., Grear, D. A., Portacci, K., Miller, R. S., Lombard, J. E., and Webb, C. T. 2013. A national-scale picture of U.S. cattle movements obtained from Interstate Certificate of Veterinary Inspection data. *Preventive Veterinary Medicine* 112:318–329.

Buluç, A., Meyerhenke, H., Safro, I., Sanders, P., and Schulz, C. 2013. Recent Advances in Graph Partitioning. *arXiv* pp. 1–36.

Charleston, B., Bankowski, B. M., Gubbins, S., Chase-Topping, M. E., Schley, D., Howey, R., Barnett, P. V., Gibson, D., Juleff, N. D., and

Woolhouse, M. E. J. 2011. Relationship between clinical signs and transmission of an infectious disease and the implications for control. *Science (New York, N.Y.)* 332:726–9.

Christley, R. M., Pinchbeck, G. L., Bowers, R. G., Clancy, D., French, N. P., Bennett, R., and Turner, J. 2005. Infection in social networks: Using network analysis to identify high-risk individuals. *American Journal of Epidemiology* 162:1024–1031.

Clauset, A., Newman, M. E. J., and Moore, C. 2004. Finding community structure in very large networks. *Physical Review E* 70:066111.

Clauset, A., Shalizi, C. R., and Newman, M. E. J. 2009. Power-law Distributions In Empirical Data. *SIAM Review* 51:661–73.

Danon, L. and Díaz-Guilera, A. 2005. Comparing community structure identification. *Journal of Statistical Mechanics: Theory and Experiment* 09008.

Danon, L., Ford, A. P., House, T., Jewell, C. P., Keeling, M. J., Roberts, G. O., Ross, J. V., and Vernon, M. C. 2011. Networks and the Epidemiology of Infectious Disease. *Interdisciplinary Perspectives on Infectious Diseases* 2011:1–28.

Danon, L., House, T., and Keeling, M. J. 2009. The role of routine versus random movements on the spread of disease in Great Britain. *Epidemics* 1:250–8.

Deardon, R., Brooks, S. P., Grenfell, B. T., Keeling, M. J., Tildesley, M. J., Savill, N. J., Shaw, D. J., and Woolhouse, M. E. J. 2010. Inference for individual-level models of infectious diseases in large populations. *Statistica Sinica* 20:239–261.

DEFRA 2012. Livestock at farm shows and markets: welfare regulations.

Dietz, K. 1993. The estimation of the basic reproduction number for infectious diseases. *Statistical Methods in Medical Research* 2:23–41.

Doel, T. R. 2003. FMD vaccines. *Virus Research* 91:81–99.

Doel, T. R. 2005. Natural and vaccine induced immunity to FMD, pp. 103–131. *In* B. W. J. Mahy (ed.), Foot-and-Mouth Disease Virus, volume 288. Springer Berlin Heidelberg.

Dubé, C., Ribble, C., Kelton, D., and McNab, B. 2009. A review of network analysis terminology and its application to foot-and-mouth disease modelling and policy development. *Transboundary and Emerging Diseases* 56:73–85.

Dubé, C., Stevenson, M. A., Garner, M. G., Sanson, R. L., Corso, B. A., Harvey, N., Griffin, J., Wilesmith, J. W., and Estrada, C. 2007. A comparison of predictions made by three simulation models of foot-and-mouth disease. *New Zealand veterinary journal* 55:280–8.

Dutta, B. L., Ezanno, P., and Vergu, E. 2014. Characteristics of the spatio-temporal network of cattle movements in France over a 5-year period. *Preventive Veterinary Medicine* 117:79–94.

Ensoy, C., Aerts, M., Welby, S., Van der Stede, Y., and Faes, C. 2013. A dynamic spatio-temporal model to investigate the effect of cattle movements on the spread of bluetongue BTV-8 in Belgium. *PloS one* 8:e78591.

Expert, P., Evans, T. S., Blondel, V. D., and Lambiotte, R. 2011. Uncovering space-independent communities in spatial networks. *Proceedings of the National Academy of Sciences of the United States of America* 108:7663–8.

FAO and OIE 2012. The Global Foot and Mouth Disease Control Strategy. Technical report, OIE.

Ferguson, N. M., Donnelly, C. A., and Anderson, R. M. 2001a. The foot-and-mouth epidemic in Great Britain: pattern of spread and impact of interventions. *Science (New York, N.Y.)* 292:1155–60.

Ferguson, N. M., Donnelly, C. A., and Anderson, R. M. 2001b. Transmission intensity and impact of control policies on the foot and mouth epidemic in Great Britain. *Nature* 413:542–8.

Fine, P. E. M. 1993. Herd immunity: history, theory, practice. *Epidemiologic reviews* 15:265–302.

Fortunato, S. 2010. Community detection in graphs. *Physics Reports* 486:75–174.

Foster, J. G., Foster, D. V., Grassberger, P., and Paczuski, M. 2010. Edge direction and the structure of networks. *Proceedings of the National Academy of Sciences of the United States of America* 107:10815–20.

Fumanelli, L., Ajelli, M., Manfredi, P., Vespignani, A., and Merler, S. 2012. Inferring the structure of social contacts from demographic data in the analysis of infectious diseases spread. *PLoS computational ...* 8.

Garner, M. G. and Beckett, S. D. 2005. Modelling the spread of foot-and-mouth disease in Australia. *Australian Veterinary Journal* 83:758–766.

Gibbens, J. C., Wilesmith, J. W., Sharpe, C. E., Mansley, L. M., Michalopoulou, E., Ryan, J. B. M., and Hudson, M. 2001. Descriptive epidemiology of the 2001 foot-and-mouth disease epidemic in Great Britain: the first five months. *Veterinary Record* 149:729–743.

Gilbert, M., Aktas, S., Mohammed, H., Roeder, P., Sumption, K., Tufan, M., and Slingenbergh, J. 2005. Patterns of spread and persistence of foot-and-mouth disease types A, O and Asia-1 in Turkey: a meta-population approach. *Epidemiology and Infection* 133:537–545.

Gonzales, J. L., Barrientos, M. A., Quiroga, J. L., Ardaya, D., Daza, O., Martinez, C., Orozco, C., Crowther, J., and Paton, D. J. 2014. Within herd transmission and evaluation of the performance of clinical and serological diagnosis of foot-and-mouth disease in partially immune cattle herds. *Vaccine* 32:6193–6198.

González, M. C., Hidalgo, C. A., and Barabási, A.-L. 2008. Understanding individual human mobility patterns. *Nature* 453:779–82.

Goodman, L. 1961. Snowball Sampling. *The Annals of Mathematical Statistics* 32:148–170.

Green, D. M. and Kao, R. R. 2007. Data quality of the Cattle Tracing System in Great Britain. *Veterinary Record* 161:439–443.

Green, D. M., Kiss, I. Z., and Kao, R. R. 2006. Modelling the initial spread of foot-and-mouth disease through animal movements. *Proceedings. Biological sciences / The Royal Society* 273:2729–35.

Green, D. M., Kiss, I. Z., Mitchell, A. P., and Kao, R. R. 2008. Estimates for local and movement-based transmission of bovine tuberculosis in British cattle. *Proceedings. Biological sciences / The Royal Society* 275:1001–5.

Greenwood, M. 1927. Some Epidemiological Observations on Foot-and-Mouth Disease , with Special Reference to the Recent Experience of Holland. *The Journal of Hygiene* 26:465–489.

Grubman, M. J. and Baxt, B. 2004. Foot-and-mouth disease. *Clinical microbiology reviews* 17:465–93.

Guimerà, R. and Amaral, L. A. N. 2005. Functional cartography of complex metabolic networks. *Nature* 433:895–900.

Halasa, T., Boklund, A., Stockmarr, A., Enøe, C., and Christiansen, L. E. 2014. A comparison between two simulation models for spread of foot-and-mouth disease. *PLoS ONE* 9.

Hastings, W. K. 1970. Monte Carlo sampling methods using Markov chains and their applications. 57:97–109.

Heesterbeek, J. A. 2002. A brief history of R0 and a recipe for its calculation. *Acta Biotheor* 50:189–204.

House, T. and Keeling, M. J. 2011. Epidemic prediction and control in clustered populations. *Journal of Theoretical Biology* 272:1–7.

IAEA and FAO 2007. The Use of Non-structural Proteins of Foot and Mouth Disease Virus (FMDV) to Differentiate Between Vaccinated and Infected Animals. Technical Report May, International Atomic Energy Agency, Vienna.

Jamal, S. M. and Belsham, G. J. 2013. Foot-and-mouth disease : past , present and future. *Veterinary research* 44:1–14.

Jutla, I. S., Jeub, L. G. S., and Mucha, P. J. 2014. A generalized Louvain method for community detection implemented in MATLAB.

Kao, R. R. 2002. The role of mathematical modelling in the control of the 2001 FMD epidemic in the UK. *Trends in Microbiology* 10:279–286.

Kao, R. R., Danon, L., Green, D. M., and Kiss, I. Z. 2006. Demographic structure and pathogen dynamics on the network of livestock movements in Great Britain. *Proceedings. Biological sciences / The Royal Society* 273:1999–2007.

Kao, R. R., Gravenor, M. B., Baylis, M., Bostock, C. J., Chihota, C. M., Evans, J. C., Goldmann, W., Smith, A. J. A., and McLean, A. R. 2002. The potential size and duration of an epidemic of bovine spongiform encephalopathy in British sheep. *Science (New York, N.Y.)* 295:332–335.

Kao, R. R., Green, D. M., Johnson, J., and Kiss, I. Z. 2007. Disease dynamics over very different time-scales: foot-and-mouth disease and scrapie on the network

of livestock movements in the UK. *Journal of the Royal Society, Interface / the Royal Society* 4:907–16.

Karrer, B., Levina, E., and Newman, M. E. J. 2008. Robustness of community structure in networks. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics* 77:1–9.

Keeling, M. J. 2005a. Models of foot-and-mouth disease. *Proceedings. Biological sciences / The Royal Society* 272:1195–202.

Keeling, M. J. 2005b. The implications of network structure for epidemic dynamics. *Theoretical population biology* 67:1–8.

Keeling, M. J., Danon, L., Vernon, M. C., and House, T. A. 2010. Individual identity and movement networks for disease metapopulations. *Proceedings of the National Academy of Sciences of the United States of America* 107:8866–70.

Keeling, M. J. and Eames, K. T. D. 2005. Networks and epidemic models. *Journal of the Royal Society, Interface* 2:295–307.

Keeling, M. J. and Rohani, P. 2008. Modeling Infectious Diseases in Humans and Animals. Princeton University Press, 1st edition.

Keeling, M. J., Woolhouse, M. E. J., May, R. M., Davies, G., and Grenfell, B. T. 2003. Modelling vaccination strategies against foot-and-mouth disease. *Nature* 421:136–42.

Keeling, M. J., Woolhouse, M. E. J., Shaw, D. J., Matthews, L., Chase-Topping, M., Haydon, D. T., Cornell, S. J., Kappey, J., Wilesmith, J., and Grenfell, B. T. 2001. Dynamics of the 2001 UK foot and mouth epidemic: stochastic dispersal in a heterogeneous landscape. *Science* 294:813–7.

Kendall, C., Kerr, L. R. F. S., Gondim, R. C., Werneck, G. L., Macena, R. H. M., Pontes, M. K., Johnston, L. G., Sabin, K., McFarland, W., and Carl Kendall, Ligia R.F.S. Kerr, Rogerio C. Gondim, Guilherme L. Werneck, Raimunda Hermelinda Maia Macena, Marta Kerr Pontesm Lisa G. Johnston, Keith Sabin, W. M. 2008. An Empirical Comparison of Respondent-driven Sampling, Time Location Sampling, and Snowball Sampling for Behavioral Surveliance in Men Who Have Sex with Men, Fortaleza, Brazil. *AIDS Behavior* 12:S97–S104.

Kermack, W. O. and McKendrick, A. G. 1927. A contribution to the Mathematical Theory of Epidemics. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* 115:700–721.

Kiss, I. Z., Green, D. M., and Kao, R. R. 2006. The network of sheep movements within Great Britain: Network properties and their implications for infectious disease spread. *Journal of the Royal Society, Interface / the Royal Society* 3:669–677.

Kitching, R. P., Hutber, A. M., and Thrusfield, M. V. 2005. A review of foot-and-mouth disease with special consideration for the clinical and epidemiological factors relevant to predictive modelling of the disease. *Veterinary journal (London, England : 1997)* 169:197–209.

Kitching, R. P. and Salt, J. S. 1995. The interference by maternally-derived antibody with active immunization of farm animals against foot-and-mouth disease. *The British veterinary journal* 151:379–389.

Kivelä, M., Arenas, A., Barthelemy, M., Gleeson, J. P., Moreno, Y., and Porter, M. a. 2014. Multilayer Networks. *Journal of Complex Networks* 2:203–271.

Knight-Jones, T. 2014. Field evaluation of foot-and-mouth disease vaccination in Turkey. PhD thesis, University of London.

Knight-Jones, T. J. D., Bulut, A. N., Gubbins, S., Stärk, K. D. C., Pfeiffer, D. U., Sumption, K. J., and Paton, D. J. 2014a. Retrospective evaluation of foot-and-mouth disease vaccine effectiveness in Turkey. *Vaccine* 32:1848–55.

Knight-Jones, T. J. D., Edmond, K., Gubbins, S., and Paton, D. J. 2014b. Veterinary and human vaccine evaluation methods. *Proceedings. Biological sciences / The Royal Society* 281:20132839.

Knight-Jones, T. J. D. and Rushton, J. 2013. The economic impacts of foot and mouth disease - what are they, how big are they and where do they occur? *Preventive veterinary medicine* 112:161–73.

Kolaczyk, E. D. 2009. Statistical Analysis of Network Data: Methods and Models. Springer, New York, 1 edition.

Krings, G., Calabrese, F., Ratti, C., and Blondel, V. D. 2009. Urban Gravity: a Model for Intercity Telecommunication Flows. *Journal of Statistical Mechanics: Theory and Experiment* 07003.

Leicht, E. A. and Newman, M. E. J. 2008. Community Structure in Directed Networks. *Physical Review Letters* 100:118703.

Levy, M. 2010. Scale-free human migration and the geography of social networks. *Physica A: Statistical Mechanics and its Applications* 389:4913–4917.

Lindström, T., Grear, D. A., Buhnerkempe, M., Webb, C. T., Miller, R. S., Portacci, K., and Wennergren, U. 2013. A bayesian approach for modeling cattle movements in the United States: scaling up a partially observed network. *PloS one* 8:e53432.

Mardones, F., Perez, A., Sanchez, J., Alkhamis, M., and Carpenter, T. 2010. Parameterization of the duration of infection stages of serotype O foot-and-mouth disease virus: an analytical review and meta-analysis with application to simulation models. *Veterinary research* 41:45.

Marjoram, P., Molitor, J., Plagnol, V., and Tavare, S. 2003. Markov chain Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences of the United States of America* 100:15324–8.

Martínez-López, B., Perez, A. M., and Sánchez-Vizcaíno, J. M. 2009. Social network analysis. Review of general concepts and use in preventive veterinary medicine. *Transboundary and Emerging Diseases* 56:109–120.

McKinley, T. J., Ross, J. V., Deardon, R., and Cook, A. R. 2014. Simulation-based Bayesian inference for epidemic models. *Computational Statistics & Data Analysis* 71:434–447.

Meil, M. 2007. Comparing clusteringsan information based distance. *Journal of Multivariate Analysis* 98:873–895.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. 1953. Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics* 21:1087–1092.

Morris, R. S., Wilesmith, J. W., Stern, M. W., Sanson, R. L., and Stevenson, M. A. 2001. Predictive spatial modelling of alternative control strategies for the foot-and-mouth disease epidemic in Great Britain , 2001. *Veterinary Record* 149:137–144.

Mucha, P. J., Richardson, T., Macon, K., Porter, M. A., and Onnela, J.-P. 2010. Community structure in time-dependent, multiscale, and multiplex networks. *Science (New York, N.Y.)* 328:876–8.

National Audit Office 2002. The 2001 Outbreak of Foot and Mouth Disease. Technical report, The Stationary Office, London.

Neal, P. J. and Roberts, G. O. 2004. Statistical inference and model selection for the 1861 Hagelloch measles epidemic. *Biostatistics* 5:249–261.

Newman, M. E. J. 2004. Analysis of weighted networks. *Physical Review E* 70:056131.

Newman, M. E. J. 2006. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences of the United States of America* 103:8577–82.

Newman, M. E. J. 2010. Networks: An introduction. Oxford University Press, Oxford, 1st edition.

Newman, M. E. J. and Girvan, M. 2004. Finding and evaluating community structure in networks. *Physical Review E* 69:026113.

Nickbakhsh, S., Matthews, L., Bessell, P. R., Reid, S. W. J., and Kao, R. R. 2011. Generating social network data using partially described networks: an example informing avian influenza control in the British poultry industry. *BMC veterinary research* 7:66.

Nishiura, H. and Omori, R. 2010. An epidemiological analysis of the foot-and-mouth disease epidemic in Miyazaki, Japan, 2010. *Transboundary and Emerging Diseases* 57:396–403.

Nöremark, M., Håkansson, N., Lindström, T., Wennergren, U., and Lewerin, S. S. 2009. Spatial and temporal investigations of reported movements, births and deaths of cattle and pigs in Sweden. *Acta veterinaria Scandinavica* 51:37.

Odden, M., Athreya, V., Rattan, S., and Linnell, J. D. C. 2014. Adaptable Neighbours: Movement Patterns of GPS-Collared Leopards in Human Dominated Landscapes in India. *PLoS ONE* 9:e112044.

O'Hare, A., Orton, R. J., Bessell, P. R., and Kao, R. R. 2014. Estimating epidemiological parameters for bovine tuberculosis in British cattle using a Bayesian partial-likelihood approach. *Proceedings of the Royal Society B* 281.

OIE 2012. Foot-and-mouth disease, pp. 1–29. *In* OIE (ed.), The Manual of Diagnostic Tests and Vaccines for Terrestrial Animals (Terrestrial Manual), number 1, chapter 2.1.5. Office International des Epizooties, 7 edition.

OIE 2015. FMD - Official status.

Orsel, K., de Jong, M. C. M., Bouma, A., Stegeman, J. A., and Dekker, A. 2007. The effect of vaccination on foot and mouth disease virus transmission among dairy cows. *Vaccine* 25:327–335.

Ortiz-Pelaez, A., Pfeiffer, D. U., Soares-Magalhães, R. J., and Guitian, F. J. 2006. Use of social network analysis to characterize the pattern of animal movements in the initial phases of the 2001 foot and mouth disease (FMD) epidemic in the UK. *Preventive veterinary medicine* 76:40–55.

Paton, D. J., Füssel, A.-E., Vosloo, W., Dekker, A., and De Clercq, K. 2014. The use of serosurveys following emergency vaccination, to recover the status of foot-and-mouth disease free where vaccination is not practised. *Vaccine* 32:7050–7056.

Pindolia, D. K., Garcia, A. J., Huang, Z., Smith, D. L., Alegana, V. A., Noor, A. M., Snow, R. W., and Tatem, A. J. 2013. The demographics of human and malaria movement and migration patterns in East Africa. *Malaria journal* 12:397.

Pomeroy, L. W., Bjørnstad, O. N., Kim, H., Jumbo, S. D., Abdoulkadiri, S., and Garabed, R. 2015. Serotype-Specific Transmission and Waning Immunity of Endemic Foot-and-Mouth Disease Virus in Cameroon. *Plos One* 10:e0136642.

Rautureau, S., Dufour, B., and Durand, B. 2012. Structuring the passive surveillance network improves epizootic detection and control efficacy: a simulation study on foot-and-mouth disease in France. *Transboundary and Emerging Diseaseserging diseases* 59:311–22.

Robinson, P. W., Costa, D. P., Crocker, D. E., Gallo-Reynoso, J. P., Champagne, C. D., Fowler, M. A., Goetsch, C., Goetz, K. T.,

Hassrick, J. L., Hückstädt, L. A., Kuhn, C. E., Maresh, J. L., Maxwell, S. M., McDonald, B. I., Peterson, S. H., Simmons, S. E., Teutschel, N. M., Villegas-Amtmann, S., and Yoda, K. 2012. Foraging behavior and success of a mesopelagic predator in the northeast Pacific Ocean: insights from a data-rich species, the northern elephant seal. *PloS one* 7:e36728.

Robinson, S. E. and Christley, R. M. 2007. Exploring the role of auction markets in cattle movements within Great Britain. *Preventive veterinary medicine* 81:21–37.

Robinson, S. E., Everett, M. G., and Christley, R. M. 2007. Recent network evolution increases the potential for large epidemics in the British cattle population. *Journal of the Royal Society, Interface / the Royal Society* 4:669–74.

Roeder, P. L. and Taylor, W. P. 2007. Mass vaccination and herd immunity: cattle and buffalo. *Revue scientifique et technique (International Office of Epizootics)* 26:253–263.

Ross, R. 1911. The prevention of malaria. Murray, London, 2nd edition.

Rweyemamu, M., Roeder, P., Mackay, D., Sumption, K., Brownlie, J., Leforban, Y., Valarcher, J.-F., Knowles, N. J., and Saraiva, V. 2008. Epidemiological Patterns of Foot-and-Mouth Disease Worldwide. *Transboundary and Emerging Diseases* 55:57–72.

Salathé, M. and Jones, J. H. 2010. Dynamics and control of diseases in networks with community structure. *PLoS computational biology* 6:1–11.

Savill, N. J., Shaw, D. J., Deardon, R., Tildesley, M. J., Keeling, M. J., Woolhouse, M. E. J., Brooks, S. P., and Grenfell, B. T. 2006. Topographic determinants of foot and mouth disease transmission in the UK 2001 epidemic. *BMC veterinary research* 2:3.

Schoenbaum, M. A. and Disney, W. T. 2003. Modeling alternative mitigation strategies for a hypothetical outbreak of foot-and-mouth disease in the United States. *Preventive Veterinary Medicine* 58:25–52.

Shannon, C. E. 1948. A mathematical theory of communication. *Bell System Technical Journal* 27:379–423.

Sisson, S. A., Fan, Y., and Tanaka, M. M. 2009. Correction for, Sequential Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences* 106:16889–16889.

Stevenson, M. a., Sanson, R. L., Stern, M. W., O'Leary, B. D., Sujau, M., Moles-Benfell, N., and Morris, R. S. 2013. InterSpread Plus: A spatial and stochastic simulation model of disease in animal populations. *Preventive Veterinary Medicine* 109:10–24.

Stumpf, M. P. H., Wiuf, C., and May, R. M. 2005. Subnets of scale-free networks are not scale-free: sampling properties of networks. *Proceedings of the National ...* 102:4221–4224.

Sumption, K., McLaws, M., Bartels, C., Ferrari, G., de Leeuw, P., Lubroth, J., and Domenech, J. 2012. The progressive control pathway for FMD (PCP-FMD): a tool for developing sustainable long term national and regional FMD control. *In* FAO/OIE Global Conference on Foot and Mouth Disease Control, pp. 1–65, Bangkok. EuFMD, FAO.

Szendroi, B. and Csányi, G. 2004. Polynomial epidemics and clustering in contact networks. *Proceedings of the Royal Society B: Biological Sciences* 271 Suppl:S364–S366.

Szmaragd, C., Wilson, A. J., Carpenter, S., Wood, J. L. N., Mellor, P. S., and Gubbins, S. 2009. A modeling framework to describe the transmission of bluetongue virus within and between farms in Great Britain. *PloS one* 4:e7741.

Tanaka, M. M., Francis, A. R., Luciani, F., and Sisson, S. A. 2006. Using Approximate Bayesian Computation to Estimate Tuberculosis Transmission Parameters From Genotype Data. *Genetics* 173:1511–1520.

Tildesley, M. J., Bessell, P. R., Keeling, M. J., and Woolhouse, M. E. J. 2009. The role of pre-emptive culling in the control of foot-and-mouth disease. *Proceedings. Biological sciences / The Royal Society* 276:3239–48.

Tildesley, M. J., Deardon, R., Savill, N. J., Bessell, P. R., Brooks, S. P., Woolhouse, M. E. J., Grenfell, B. T., and Keeling, M. J. 2008. Accuracy of models for the 2001 foot-and-mouth epidemic. *Proceedings. Biological sciences / The Royal Society* 275:1459–68.

Tildesley, M. J., House, T. A., Bruhn, M. C., Curry, R. J., O'Neil, M., Allpress, J. L. E., Smith, G., and Keeling, M. J. 2010. Impact of spatial clustering on disease transmission and optimal control. *Proceedings of the National Academy of Sciences of the United States of America* 107:1041–6.

Tildesley, M. J., Savill, N. J., Shaw, D. J., Deardon, R., Brooks, S. P., Woolhouse, M. E. J., Grenfell, B. T., and Keeling, M. J. 2006. Optimal reactive vaccination strategies for a foot-and-mouth outbreak in the UK. *Nature* 440:83–6.

Tildesley, M. J., Volkova, V. V., and Woolhouse, M. E. J. 2011. Potential for epidemic take-off from the primary outbreak farm via livestock movements. *BMC veterinary research* 7:76.

Tizzoni, M., Bajardi, P., Decuyper, A., Kon Kam King, G., Schneider, C. M., Blondel, V., Smoreda, Z., González, M. C., and Colizza, V. 2014. On the use of human mobility proxies for modeling epidemics. *PLoS computational biology* 10:e1003716.

Traud, A. L., Kelsic, E. D., Mucha, P. J., and Porter, M. A. 2011. Structure to Characteristics in Online Collegiate Social Networks. *SIAM Review* 53:526–543.

Vernon, M. C. 2011. Demographics of cattle movements in the United Kingdom. *BMC veterinary research* 7:31.

Vernon, M. C. and Keeling, M. J. 2009. Representing the UK's cattle herd as static and dynamic networks. *Proceedings. Biological sciences / The Royal Society* 276:469–76.

Walker, D. M., Allingham, D., Lee, H. W. J., and Small, M. 2010. Parameter inference in small world network disease models with approximate Bayesian Computational methods. *Physica A: Statistical Mechanics and its Applications* 389:540–548.

Watts, D. J. and Strogatz, S. H. 1998. Collective dynamics of 'small-world' networks. *Nature* 393:440–2.

Wesolowski, A., Buckee, C. O., Pindolia, D. K., Eagle, N., Smith, D. L., Garcia, A. J., and Tatem, A. J. 2013. The Use of Census Migration Data to Approximate Human Movement Patterns across Temporal Scales. *PLoS ONE* 8:e52971.

Yoon, H., Yoon, S.-S., Kim, Y.-J., Moon, O.-K., Wee, S.-H., Joo, Y.-S., and Kim, B. 2015. Epidemiology of the foot-and-mouth disease serotype O epidemic of November 2010 to April 2011 in the Republic of Korea. *Transboundary and Emerging Diseases* 62:252–263.

Zhang, X.-S., Chase-Topping, M. E., McKendrick, I. J., Savill, N. J., and Woolhouse, M. E. J. 2010. Spread of E. coli O157 infection among Scottish cattle farms: stochastic models and model selection. *Epidemics* 2:11–20.

Zhang, X.-S. and Woolhouse, M. E. J. 2011. Escherichia coli O157 infection on Scottish cattle farms: dynamics and control. *J R Soc Interface* 8:1051–1058.

Appendix A

Appendix to chapter 2

A.1 Raw data

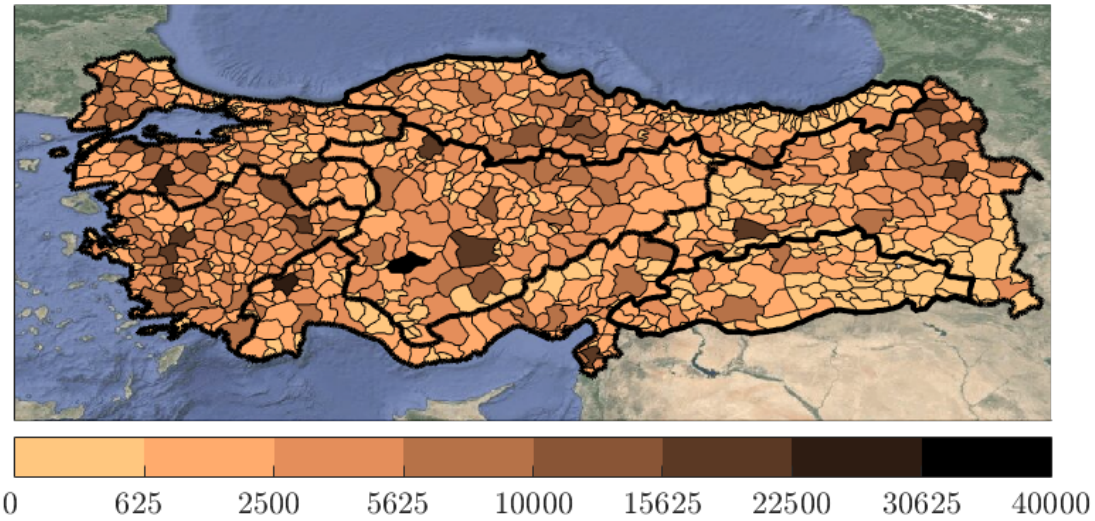

Figure A.1: The number of movements destined for each district in Turkey for the year 2010. This includes the within district movements. Note that the colour scale is non-linear so as to show more variation between areas of low movement.

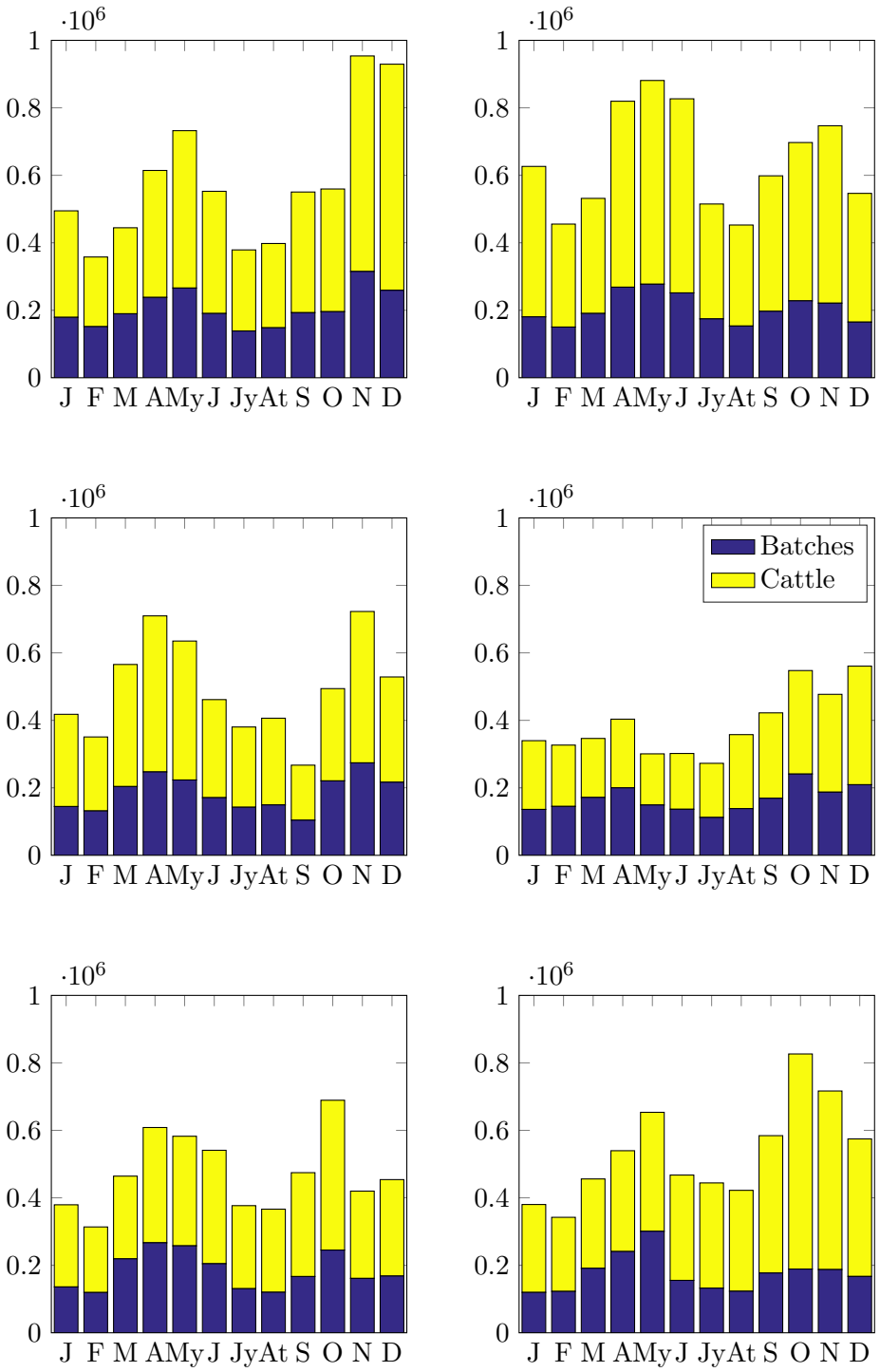

Figure A.2: Monthly number of batches and number of cattle moved from 2007-2012.

A.2 Network statistics - figures

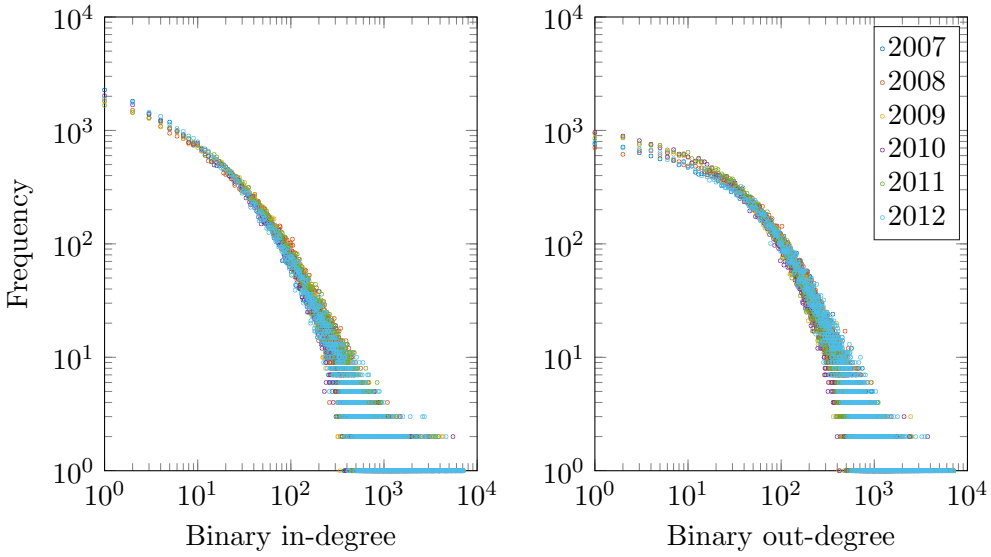

Figure A.3: Degree distributions for the binary in and out degrees from 2007 to 2012.

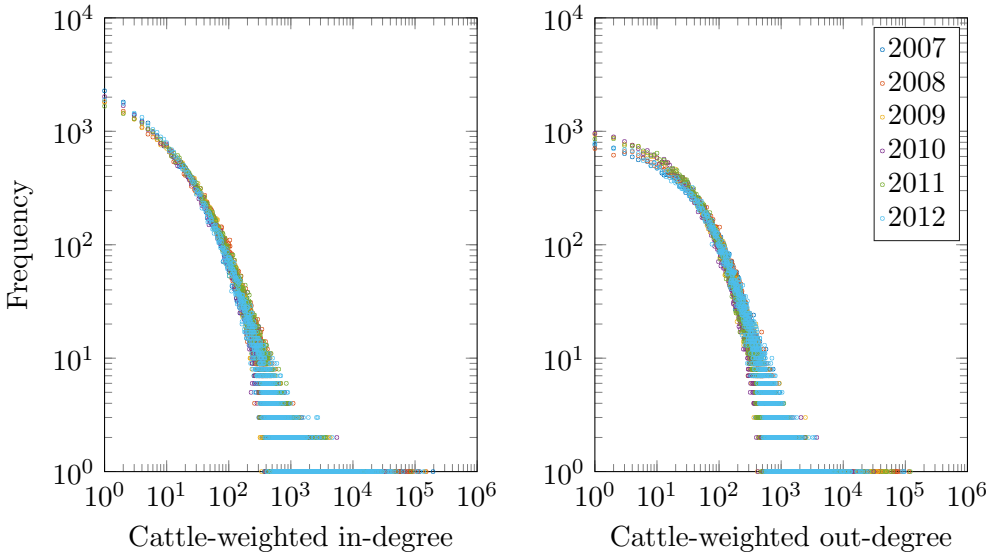

Figure A.4: Cattle weighted in and out degree distributions from 2007 to 2012.

Appendix B

Appendix to chapter 3

B.1 Equivalence of rebuilding by scaling with rebuilding by preferentially increasing edge weight

In this chapter we utilize a simple re-scaling method to rebuild partially sampled networks. An apparently more sophisticated approach would be to adapt the weighted scale free model approach of Yook *et al.*[i] to a static network where new edges are not created but edge weights are increased preferentially based upon the weights of existing edges. However, we can show that these two methods are, in fact, equivalent.

Given that the depleted network begins with some fraction α of the original number of movements N_0 and we wish to add n movements such that $N_0 = \alpha N_0 + n$ where the probability of movement from farm i to j being added is given by

$$p(i \to j, t) = \frac{a_{ij}(t)}{\sum a_{ij}(t)} = \frac{N(i \to j, t)}{\alpha N_0 + t}.$$

The weight of the edge between i and j at time $t + 1$ is then

$$\begin{aligned} N(i \to j, t+1) &= N(i \to j, t) + p(i \to j, t) \\ &= N(i \to j) + \frac{N(i \to j, t)}{\alpha N_0 + t} \end{aligned}$$

if we then enforce the initial condition that the depleted network begins with N_D movements, that is

$$N(i \to j, 0) = N_D(i \to j)$$

[i] Yook S, Jeong H, Barabási AL, and Tu Y (2001), Weighted evolving networks, *Phys Rev Lett*, 86(25):5835–5838.

this can be solved explicitly as

$$\begin{aligned} N(i \to j, n) &= N_D(i \to j) \prod_{t=1}^{n-1} \left(1 + \frac{1}{\alpha N_0 + t}\right) \\ &= N_D(i \to j) \frac{n + \alpha N_0}{1 + \alpha N_0} \\ &= N_D(i \to j) \frac{N_0}{1 + \alpha N_0} \end{aligned}$$

and assuming that $\alpha N_0 >> 1$ which it is, we obtain the the result

$$N_n(i \to j) = \frac{1}{\alpha} N_D(i \to j). \tag{B.1}$$

Therefore if we were to carry out the preferential weighting method many times the average weights would be the same as the re-scaling method we utilize in this paper.

B.2 Network statistics for RMS

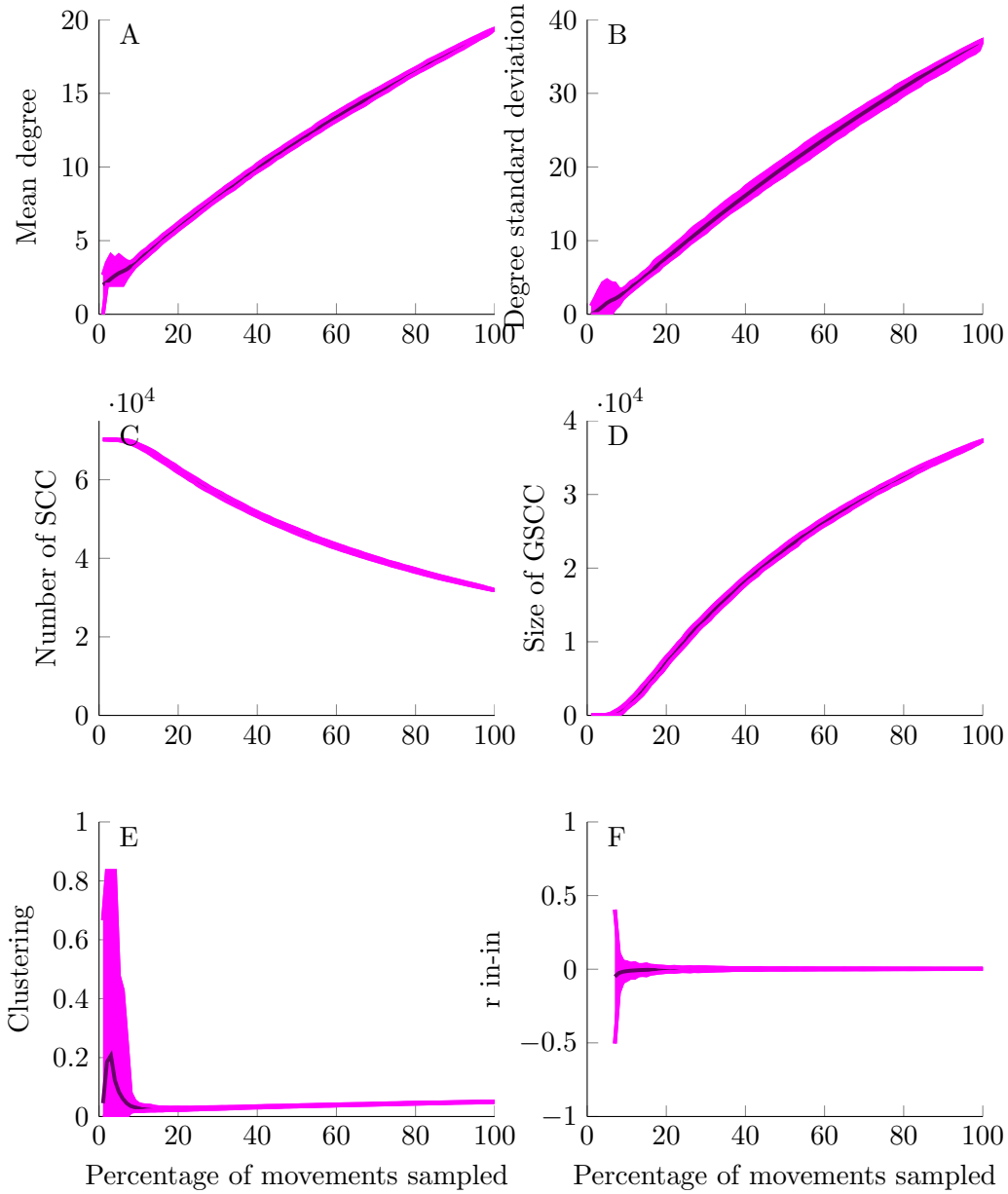

Figure B.1: Graphs showing (A) the mean degree, (B) the degree standard deviation, (C) the number of strongly connected components (D) the size of the giant strongly connected component, (E), the mean local clustering coefficient, and (F) the correlation for for the RMS method as a function of the percentage of movements sampled. These statistics are averaged over 1000 realisations with shaded confidence intervals depicting the maximum and minimum value of each statistic.

B.3 Supplementary figures for Cumbria $\beta = 1$, $T = 21$ days

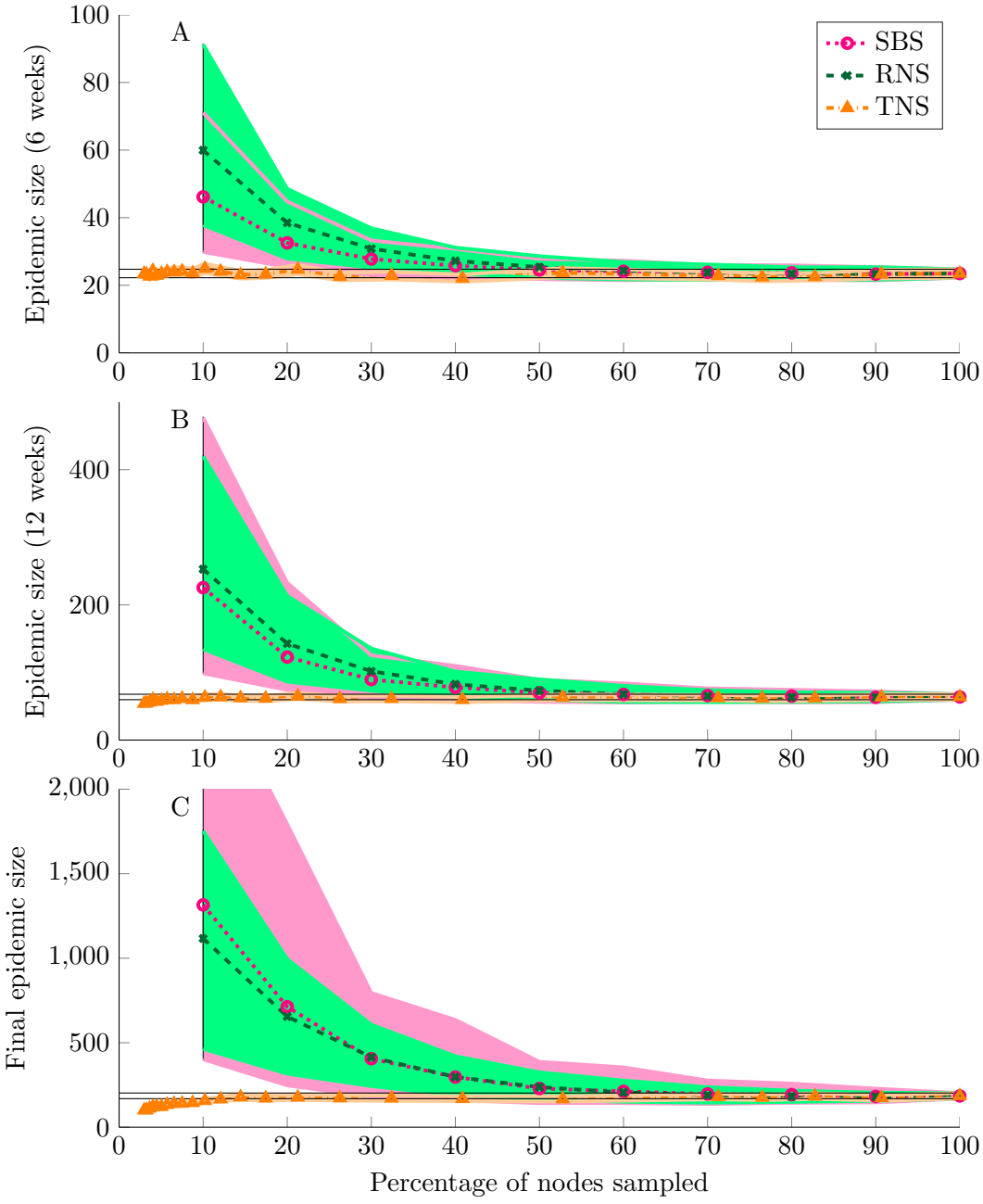

Figure B.2: Epidemic measurements for outbreaks seeded in Cumbria using the RNS (crosses) method, SBS (circles) method and the TNS (triangles) method with shaded 95% confidence intervals, when $\beta = 1$ and $T = 21$ days for (a) 6 weeks, (b) 12 weeks and (c) the whole epidemic. In this plot, the RNS and SBS methods are scaled as in equation (3.1).

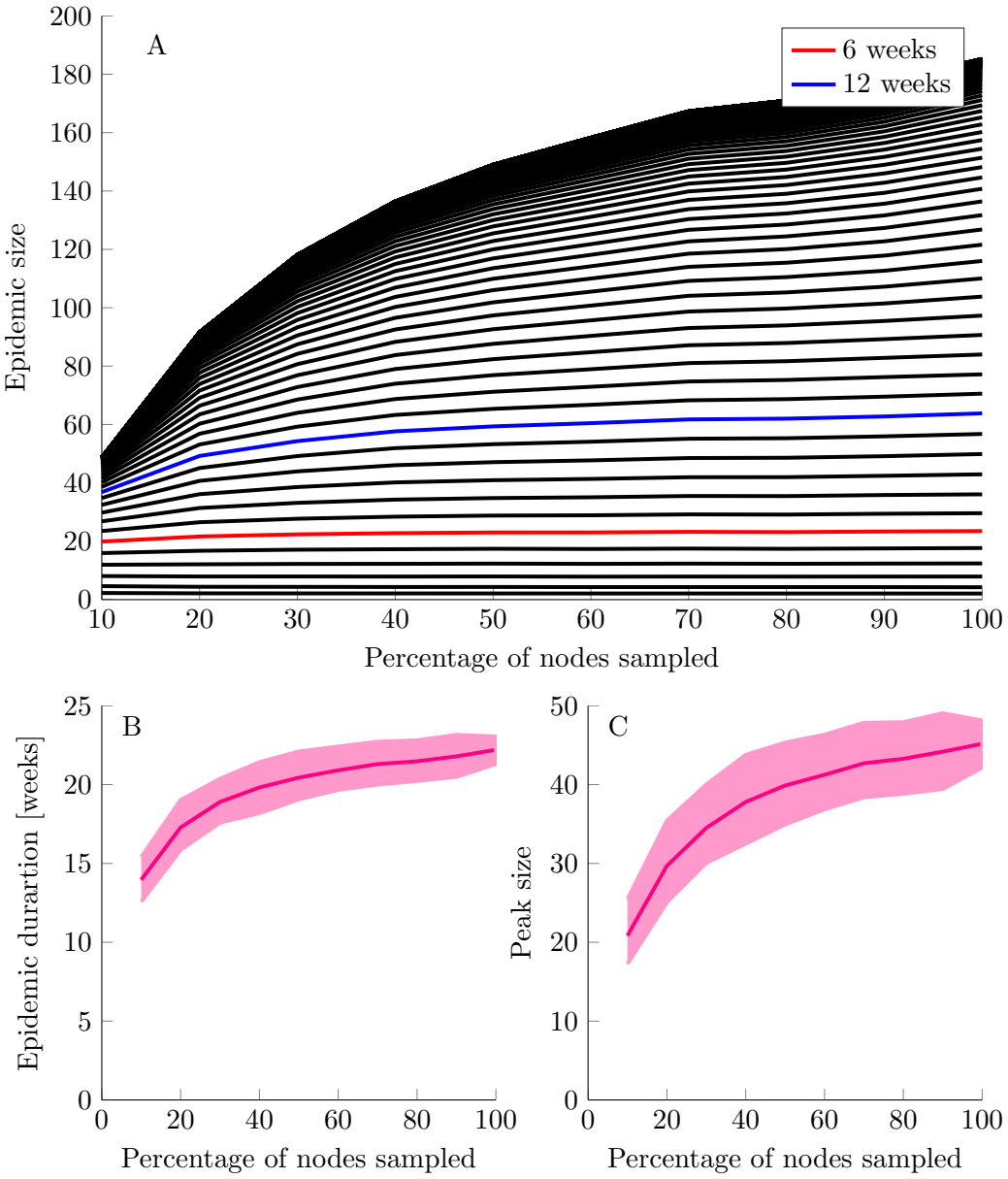

Figure B.3: Epidemic measurements for outbreaks seeded in Cumbria using the RMS method with $\beta = 1$ and $T = 21$ days. Plot a) shows the mean epidemic size in weekly intervals with ascending contours representing the epidemic size progression through time in weeks, the size after 6 (red) and 12 (blue) weeks are highlighted for reference. Plot b) shows the mean epidemic duration in weeks with 95% confidence intervals. Plot c) shows the mean peak size of the epidemic with 95% confidence intervals.

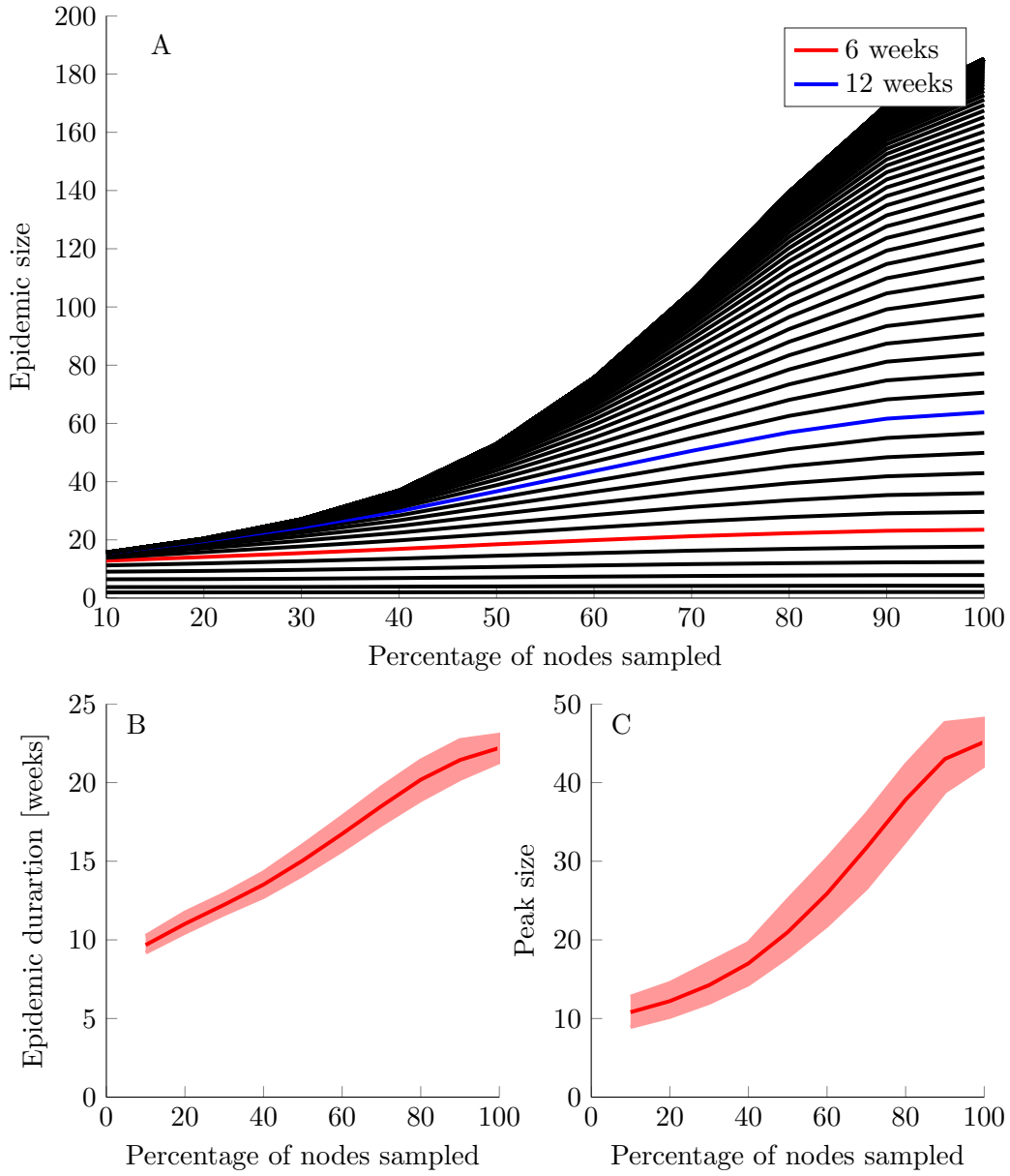

Figure B.4: Epidemic measurements for outbreaks seeded in Cumbria using the RNS method with $\beta = 1$ and $T = 21$ days. Plot a) shows the mean epidemic size in weekly intervals with ascending contours representing the epidemic size progression through time in weeks, the size after 6 (red) and 12 (blue) weeks are highlighted for reference. Plot b) shows the mean epidemic duration in weeks with 95% confidence intervals. Plot c) shows the mean peak size of the epidemic with 95% confidence intervals.

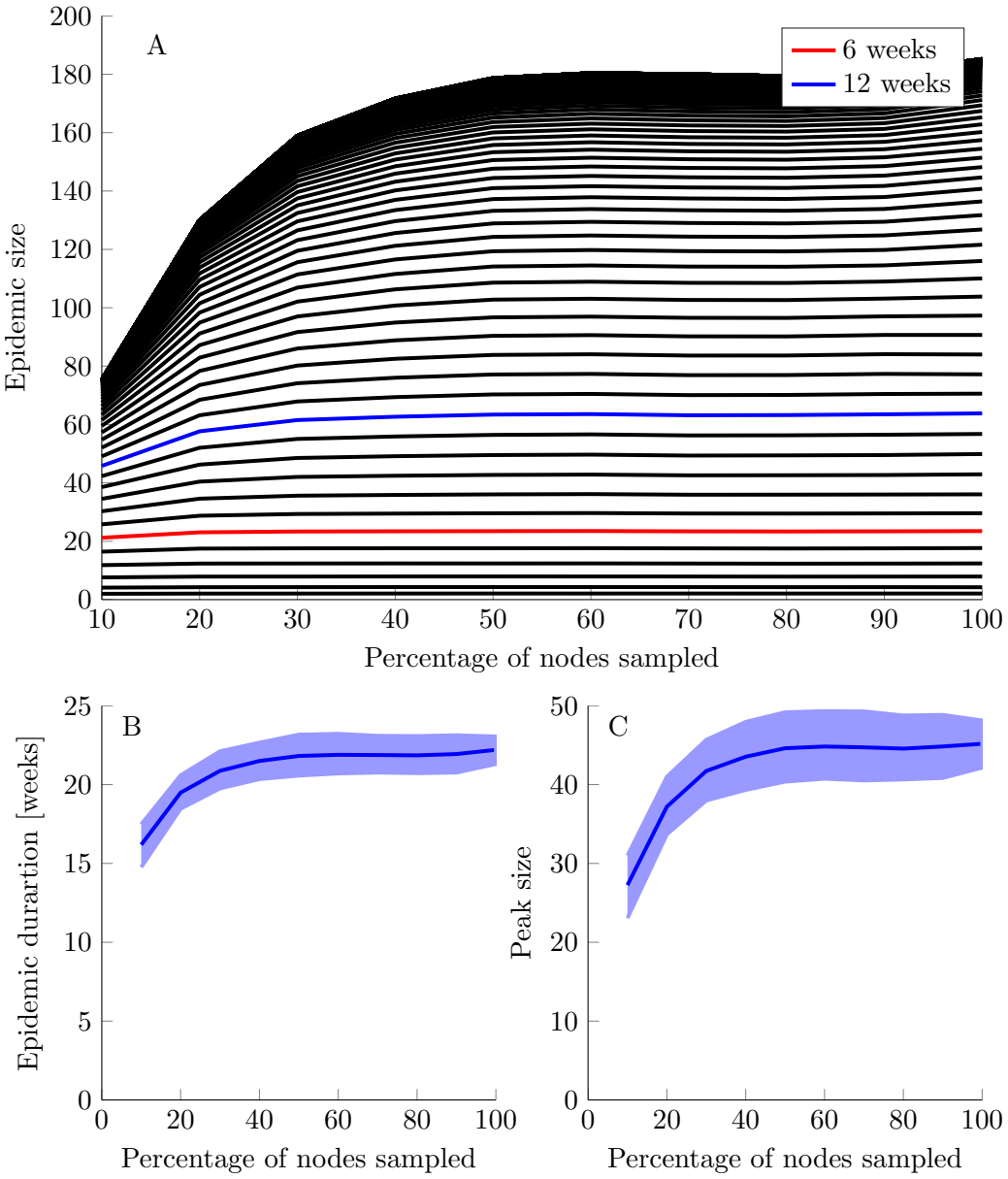

Figure B.5: Epidemic measurements for outbreaks seeded in Cumbria using the SBS method with $\beta = 1$ and $T = 21$ days. Plot a) shows the mean epidemic size in weekly intervals with ascending contours representing the epidemic size progression through time in weeks, the size after 6 (red) and 12 (blue) weeks are highlighted for reference. Plot b) shows the mean epidemic duration in weeks with 95% confidence intervals. Plot c) shows the mean peak size of the epidemic with 95% confidence intervals.

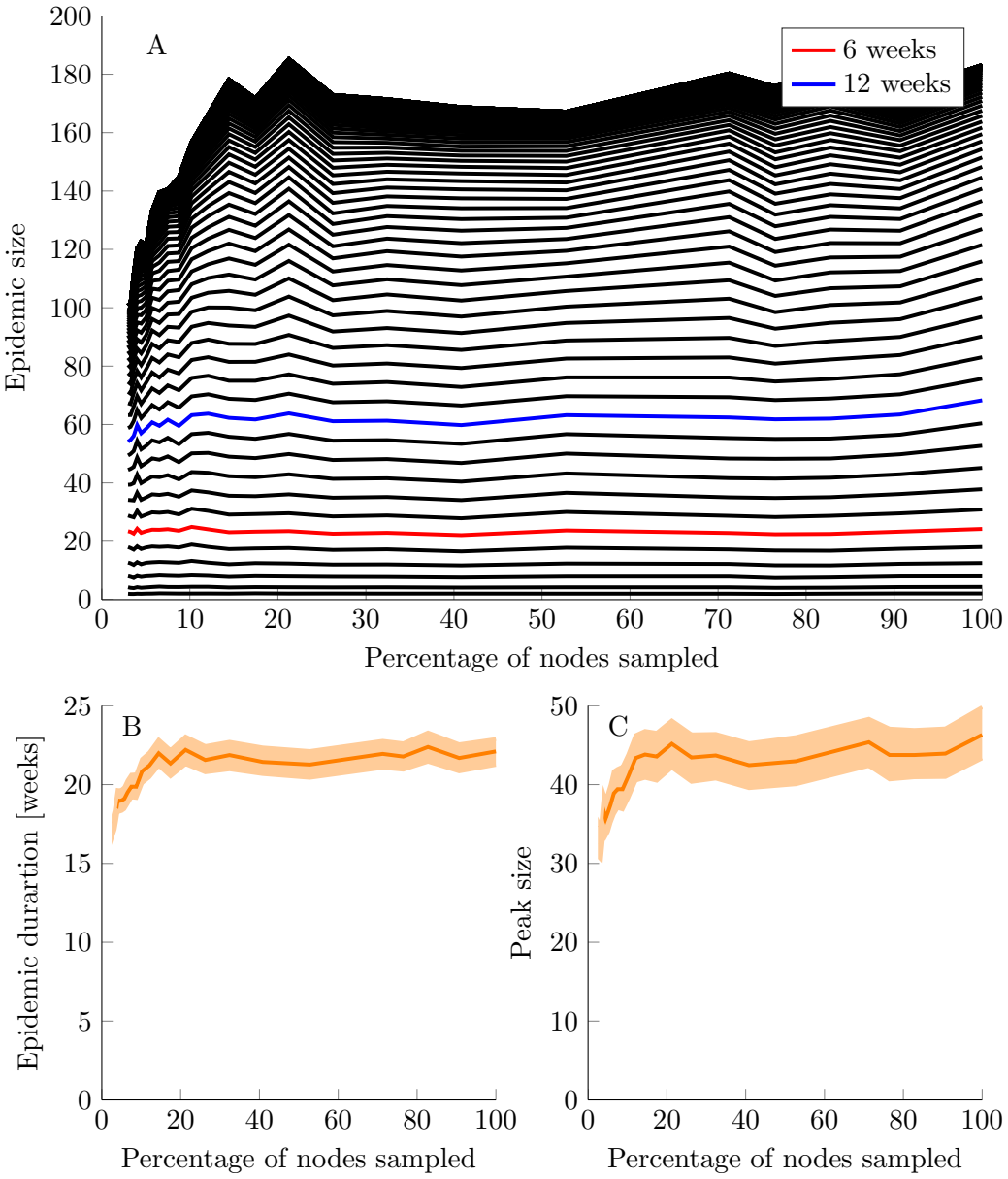

Figure B.6: Epidemic measurements for outbreaks seeded in Cumbria using the TNS method with $\beta = 1$ and $T = 21$ days. Plot a) shows the mean epidemic size in weekly intervals with ascending contours representing the epidemic size progression through time in weeks, the size after 6 (red) and 12 (blue) weeks are highlighted for reference. Plot b) shows the mean epidemic duration in weeks with 95% confidence intervals. Plot c) shows the mean peak size of the epidemic with 95% confidence intervals.

B.4 Supplementary figures for Cumbria with a range of parameters

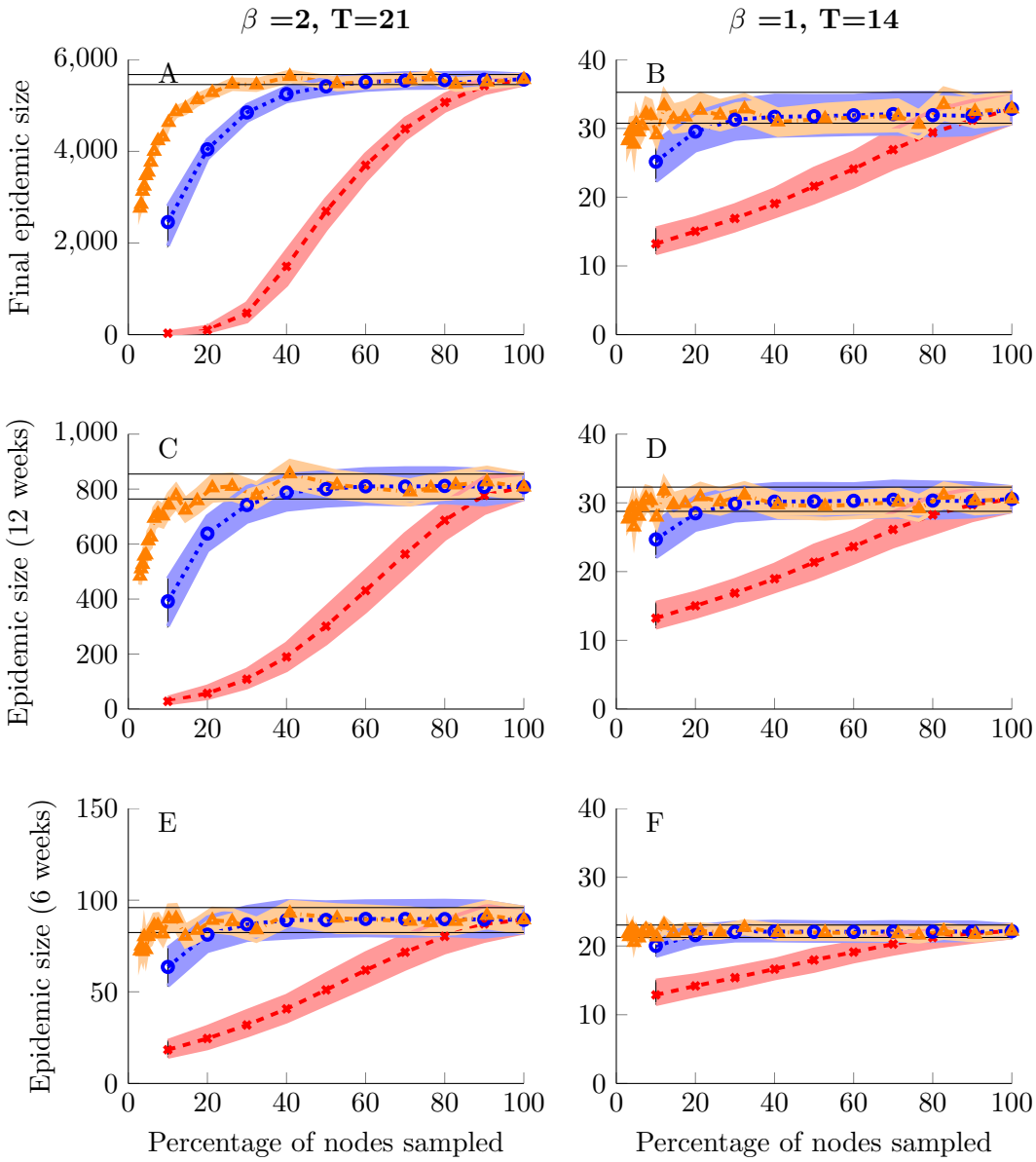

Figure B.7: Epidemic measurements for outbreaks seeded in Cumbria using the RNS (crosses), SBS (circles) and TNS (triangles) methods with $\beta = 2$ and $T = 7$ days on the left and $\beta = 1$ and $T = 14$ days on the right. Plots A and B shows the final epidemic size, plot C and D the epidemic size after 12 weeks and plot E and F the epidemic size after 6 weeks. Shaded 95% confidence intervals are included as are black reference lines indicating the 95% confidence intervals for simulations run on the complete network.

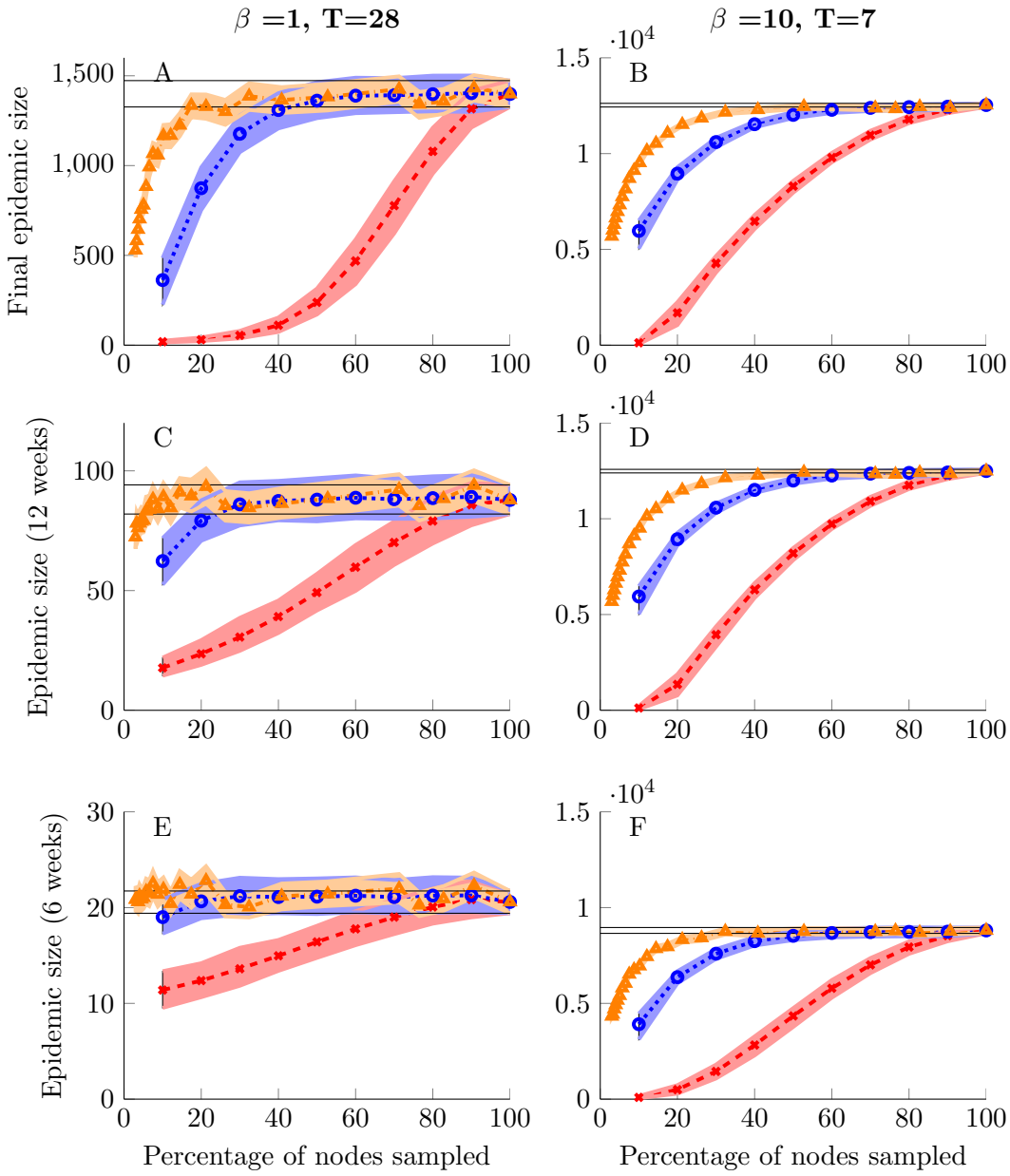

Figure B.8: Epidemic measurements for outbreaks seeded in Cumbria using the RNS (crosses), SBS (circles) and TNS (triangles) methods with $\beta = 1$ and $T = 28$ days on the left and $\beta = 10$ and $T = 07$ days on the right. Plots A and B shows the final epidemic size, plot C and D the epidemic size after 12 weeks and plot E and F the epidemic size after 6 weeks. Shaded 95% confidence intervals are included as are black reference lines indicating the 95% confidence intervals for simulations run on the complete network.

B.5 Supplementary figures for epidemics seeded in Aberdeenshire and Devon

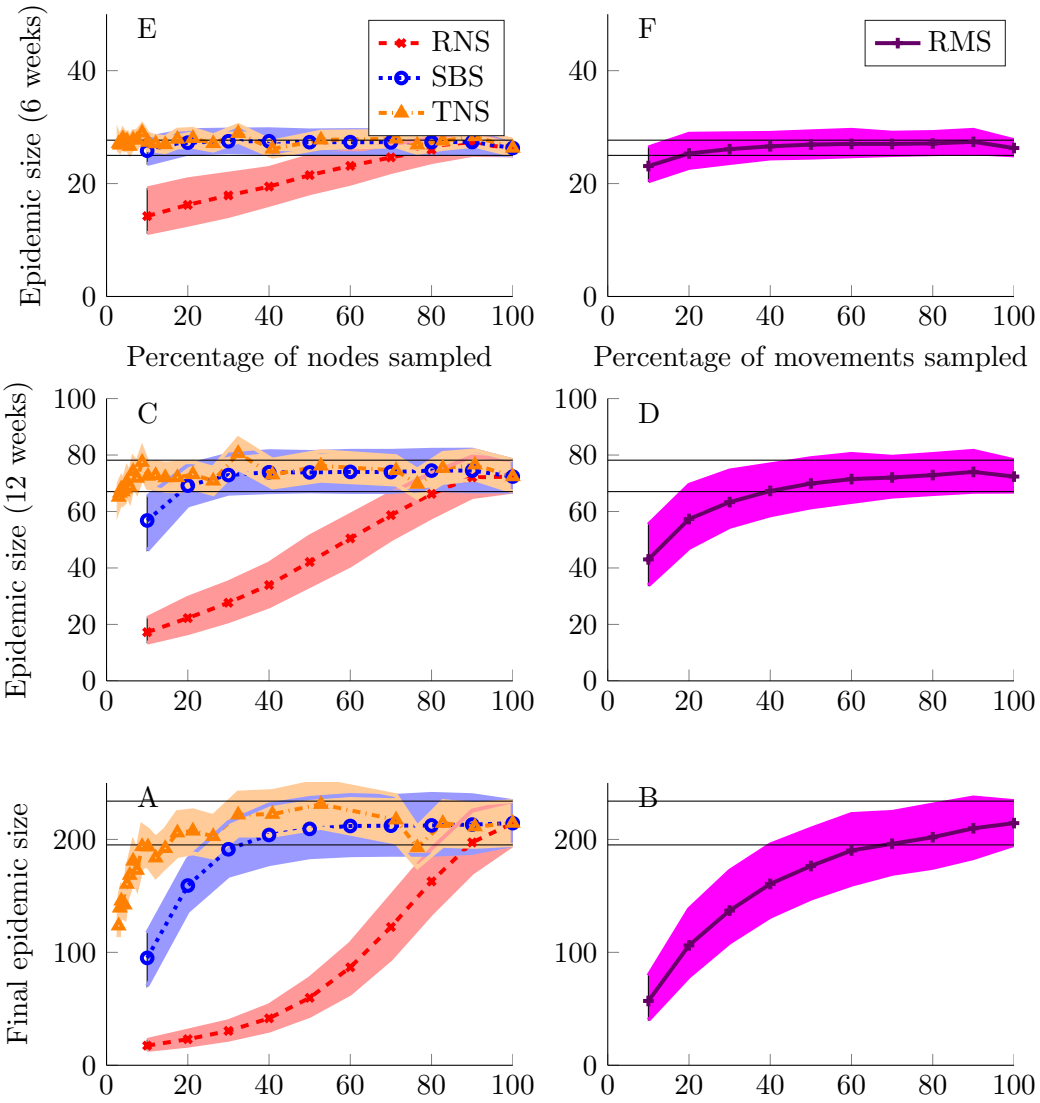

Figure B.9: Graphs (a) to (c) compare the epidemic size for outbreaks seeded in Devon on networks generated by RNS (crosses), SBS (circles) and TNS (triangles) as a function of nodes sampled with shaded 95% confidence intervals for (a) 6 weeks, (b) 12 weeks and (c) the full epidemic. The solid black lines represents the 95% confidence intervals on the average simulation for the original network. Graphs (d) to (f) show the same results for the RMS method for (d) 6 weeks, (e) 12 weeks and (f) the full epidemic.

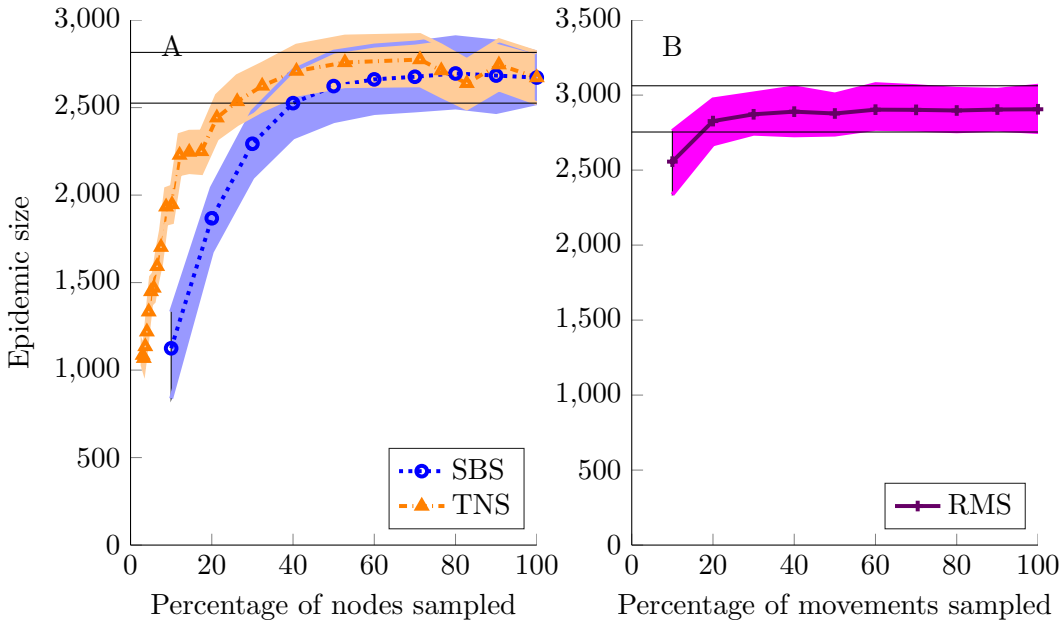

Figure B.10: Mean epidemic size for outbreaks seeded in Devon, with within-market transmission incorporated into the model when $\beta = 1$ and $T = 21$, for (a) the TNS (triangles) and the SBS (circles) methods and (b) the RMS (crosses) method. Shaded 95% confidence intervals are shown in each figure for each sampling method.

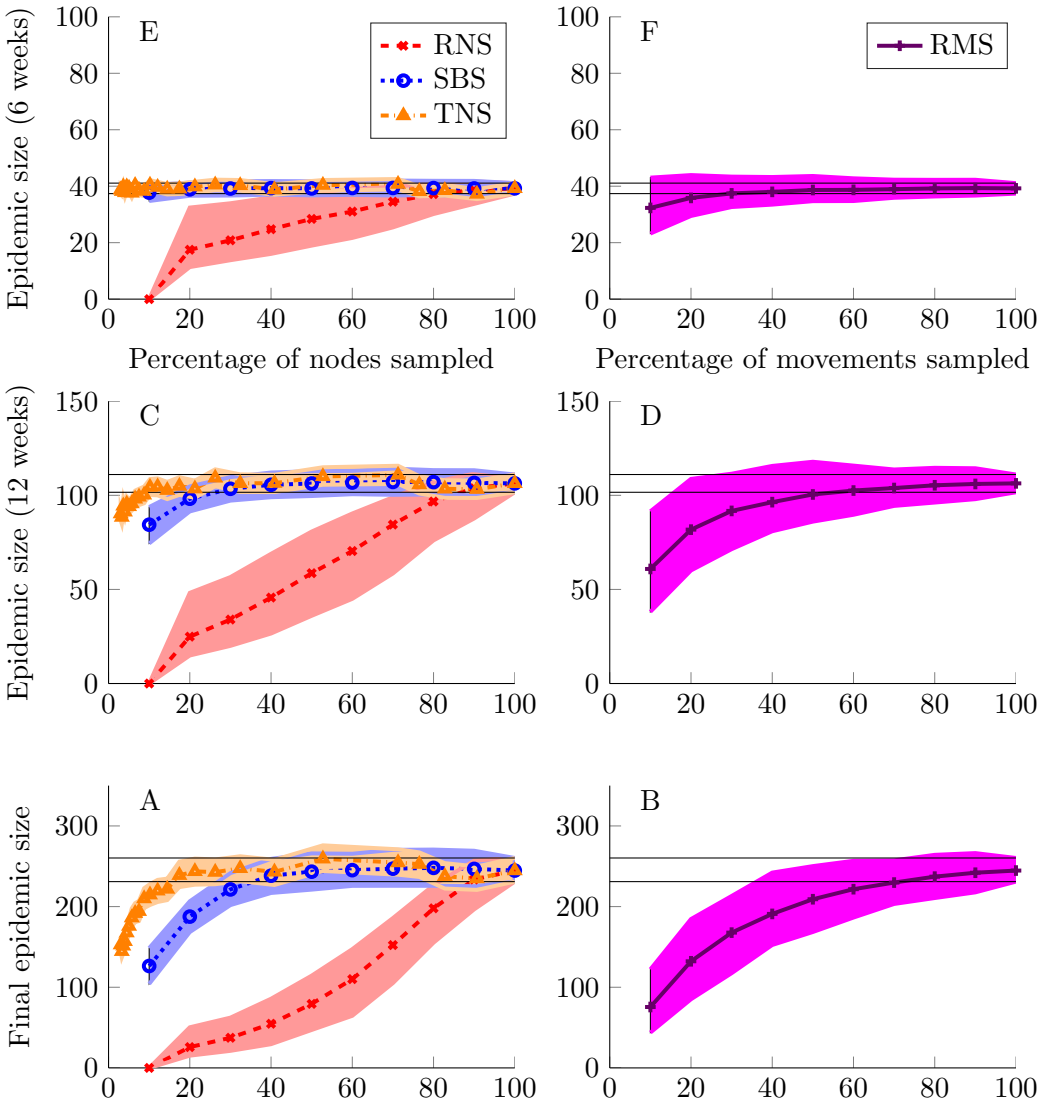

Figure B.11: Graphs (a) to (c) compare the epidemic size for outbreaks seeded in Aberdeen on networks generated by RNS (crosses), SBS (circles) and TNS (triangles) as a function of nodes sampled with shaded 95% confidence intervals for (a) 6 weeks, (b) 12 weeks and (c) the full epidemic. The solid black lines represents the 95% confidence intervals on the average simulation for the original network. Graphs (d) to (f) show the same results for the RMS method for (d) 6 weeks, (e) 12 weeks and (f) the full epidemic.

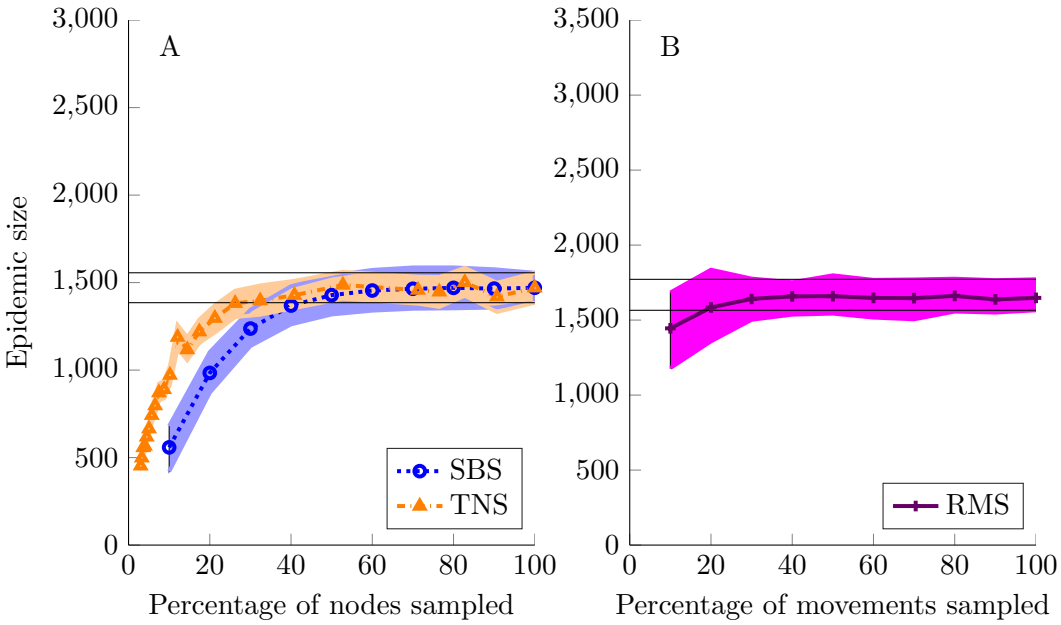

Figure B.12: Mean epidemic size for outbreaks seeded in Aberdeen, with within-market transmission incorporated into the model when $\beta = 1$ and $T = 21$, for (a) the TNS (triangles) and the SBS (circles) methods and (b) the RMS (crosses) method. Shaded 95% confidence intervals are shown in each figure for each sampling method.

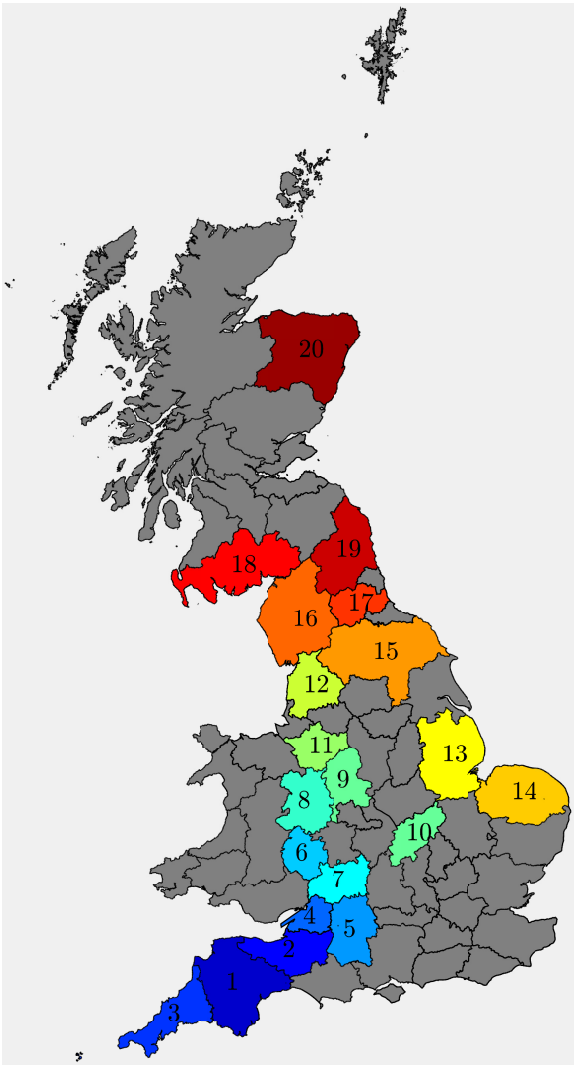

Figure B.13: Map showing the 20 counties with the largest mean number of infected farms after 12 weeks when epidemics are seeded in Devon and markets are not explicitly included. The mean size for each county is given in figure B.14

.

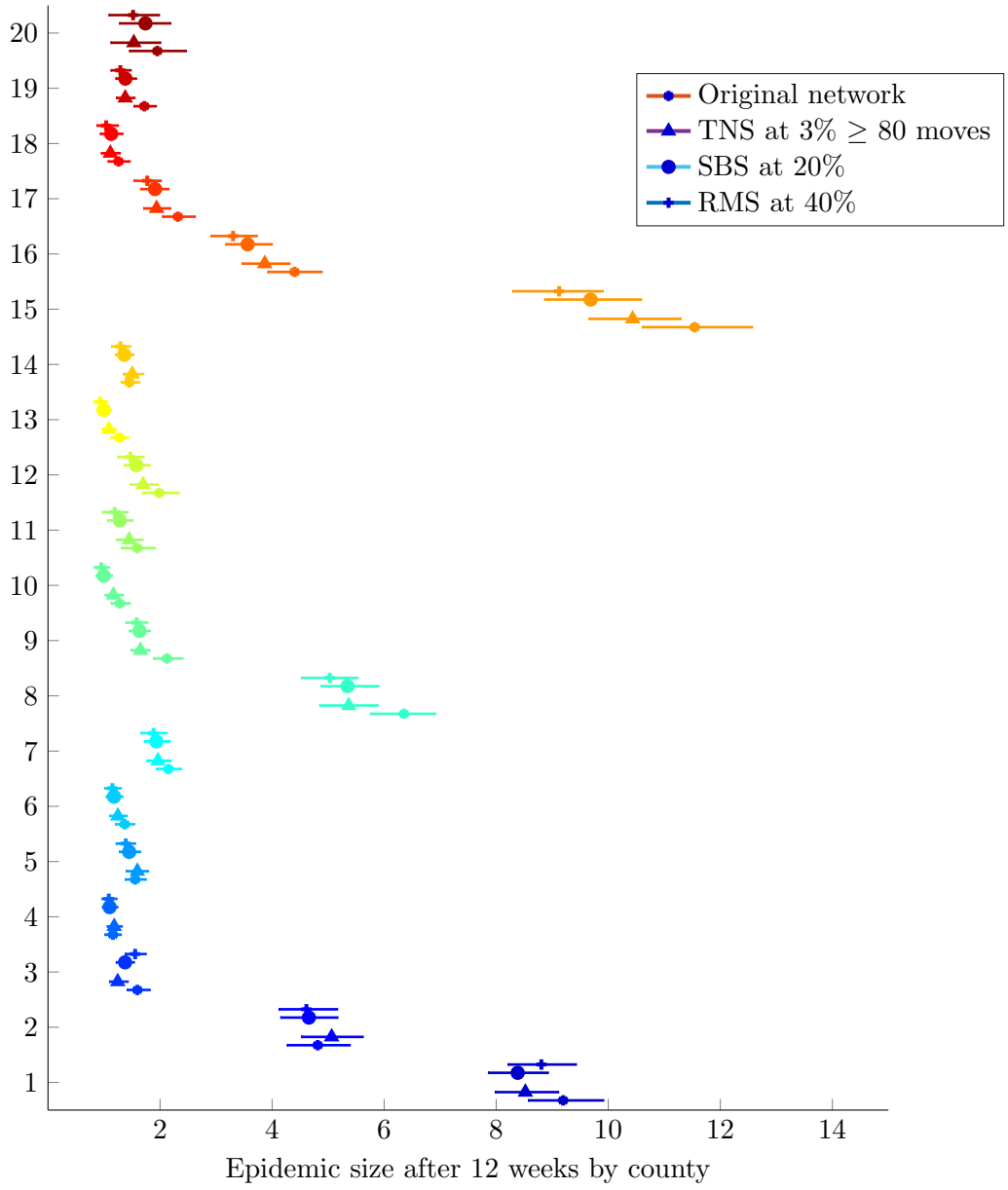

Figure B.14: The average epidemic size for the original network (stars) random movement sampling (RMS) with 50% of sampled movements (crosses), snowball sampling with 30% of nodes (circles) and targeted node sampling (TNS), sampling nodes with more than 50 movements (triangles) for the 20 most infected counties when epidemics are seeded in Devon. Counties are ordered in terms of the proximity of their centroids from Devon with labels corresponding to the map in figure B.13.

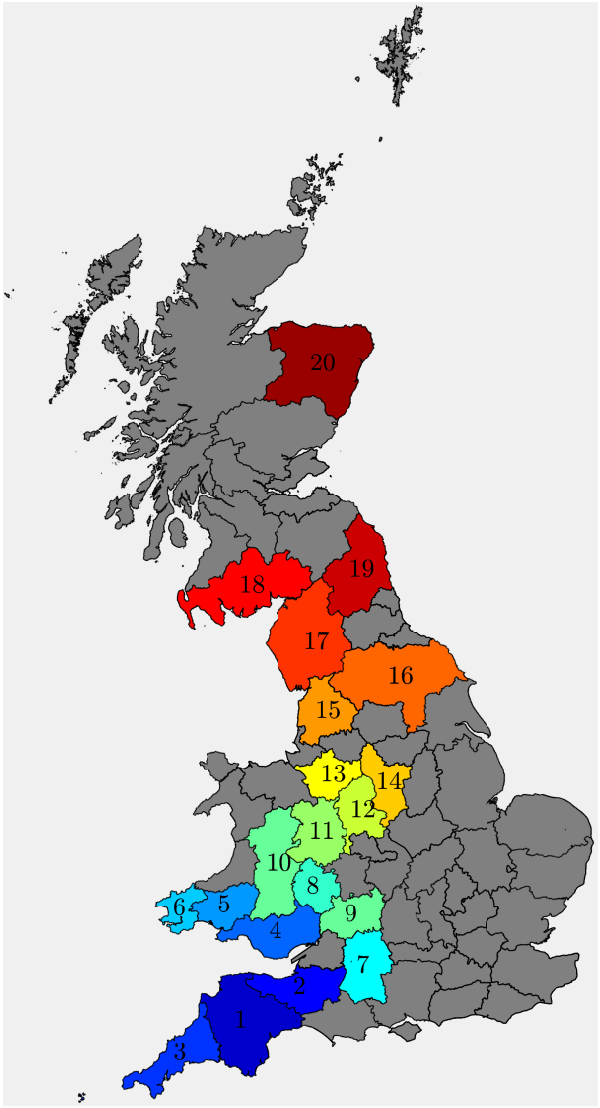

Figure B.15: Map showing the 20 counties with the largest mean number of infected farms after one infectious period (21 days) when epidemics are seeded in Devon and markets are explicitly included. The mean size for each county is given in figure B.16

.

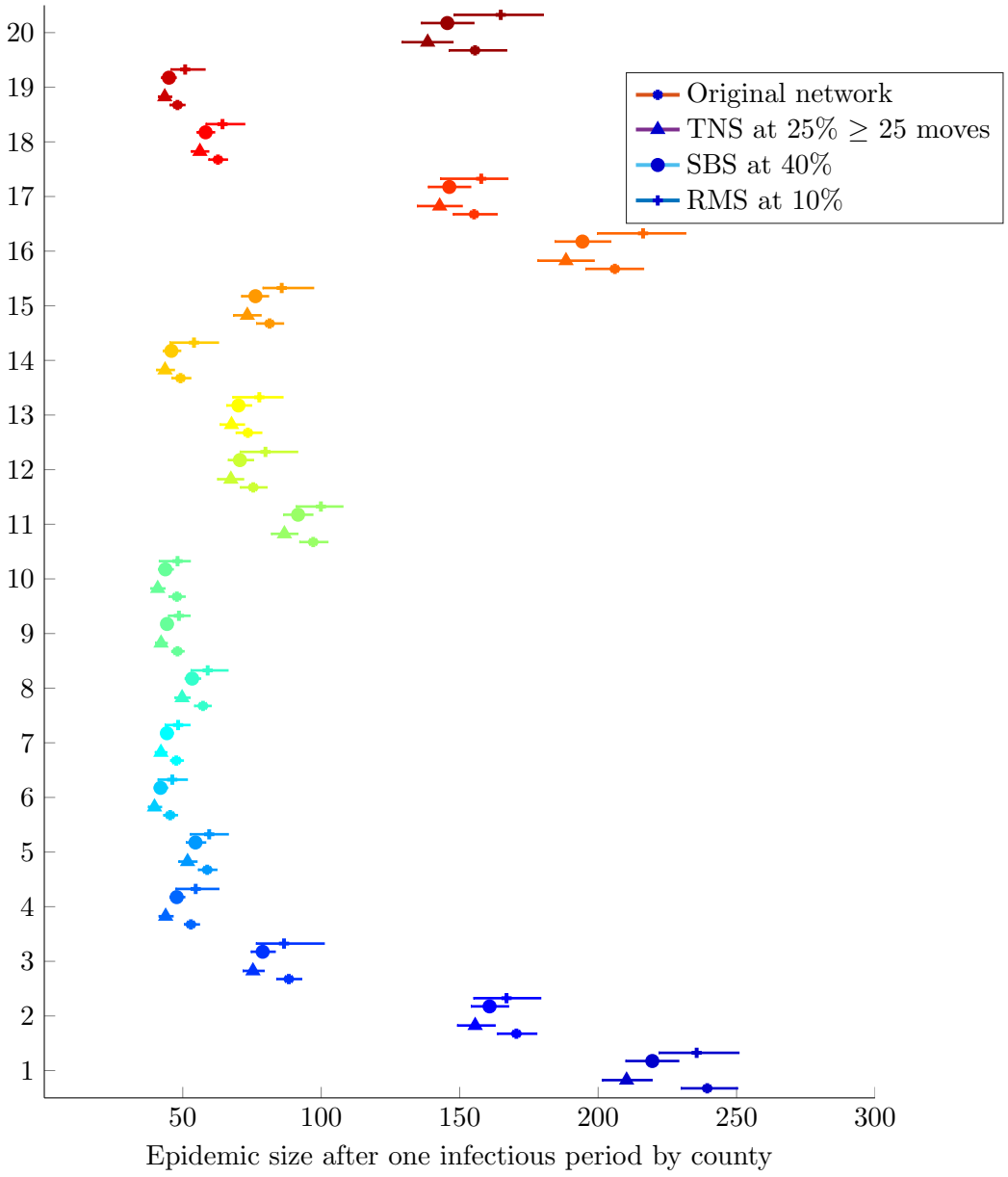

Figure B.16: The average epidemic size for the original network (stars) random movement sampling (RMS) with 10% of sampled movements (crosses), snowball sampling with 40% of nodes (circles) and targeted node sampling (TNS), sampling nodes with more than 25 movements (triangles) for the 20 most infected counties when epidemics are seeded in Aberdeen. Counties are ordered in terms of the proximity of their centroids from Aberdeen with labels corresponding to the map in figure B.15.

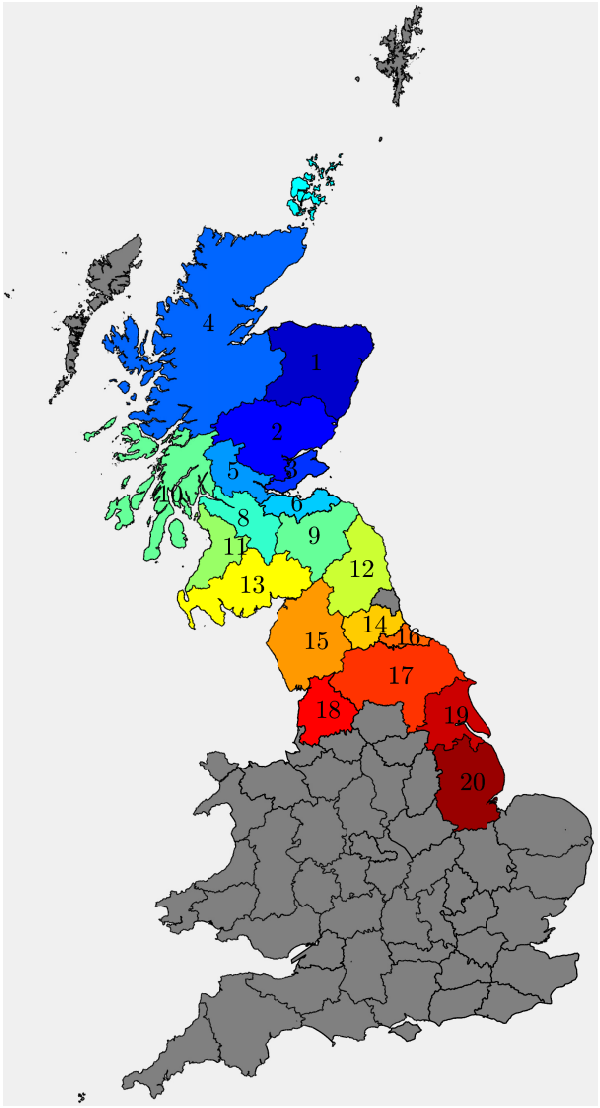

Figure B.17: Map showing the 20 counties with the largest mean number of infected farms after 12 weeks when epidemics are seeded in Aberdeen and markets are not explicitly included. The mean size for each county is given in figure B.18

.

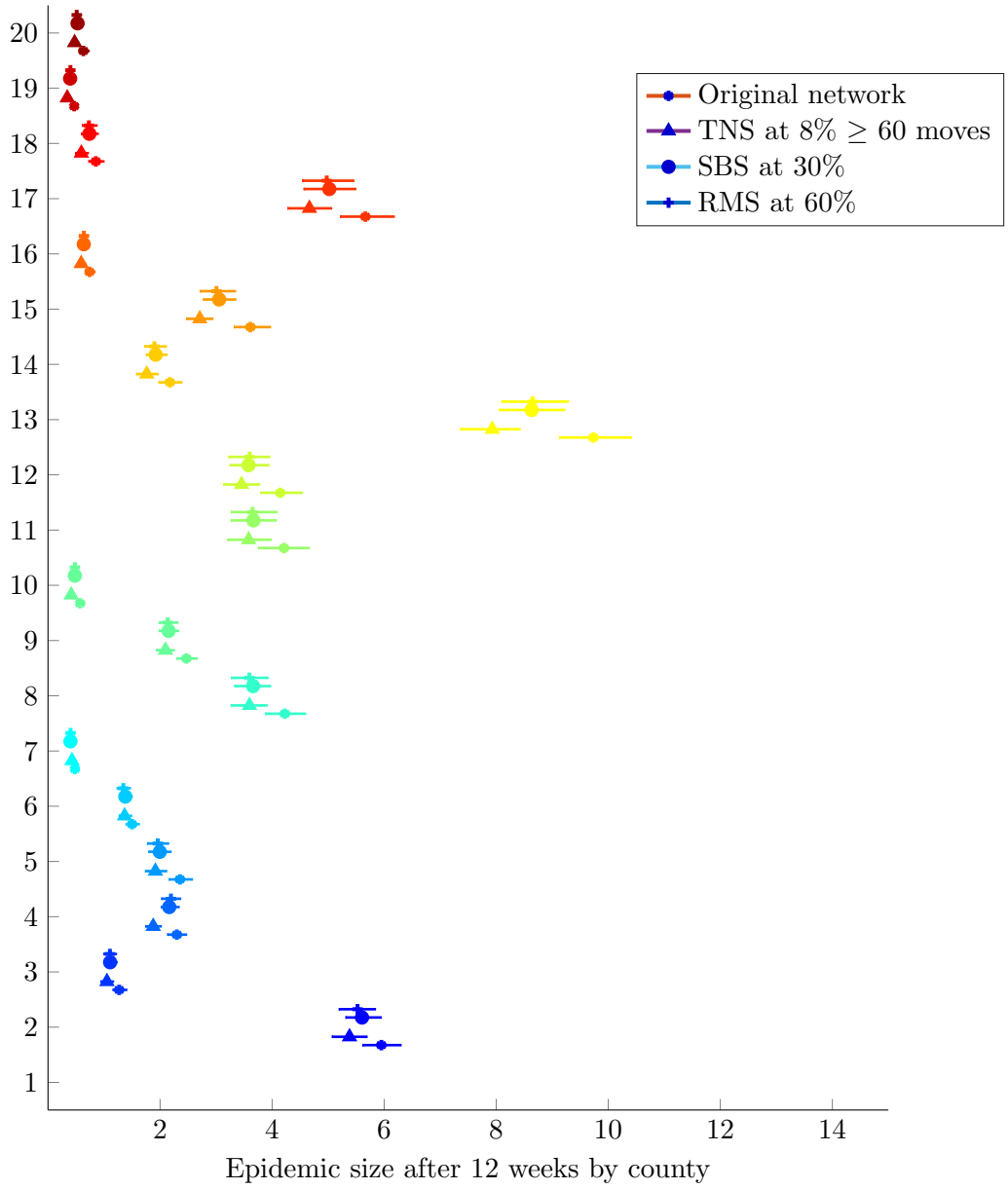

Figure B.18: The average epidemic size for the original network (stars) random movement sampling (RMS) with 50% of sampled movements (crosses), snowball sampling with 30% of nodes (circles) and targeted node sampling (TNS), sampling nodes with more than 50 movements (triangles) for the 20 most infected counties when epidemics are seeded in Aberdeen. Counties are ordered in terms of the proximity of their centroids from Aberdeen with labels corresponding to the map in figure B.13.

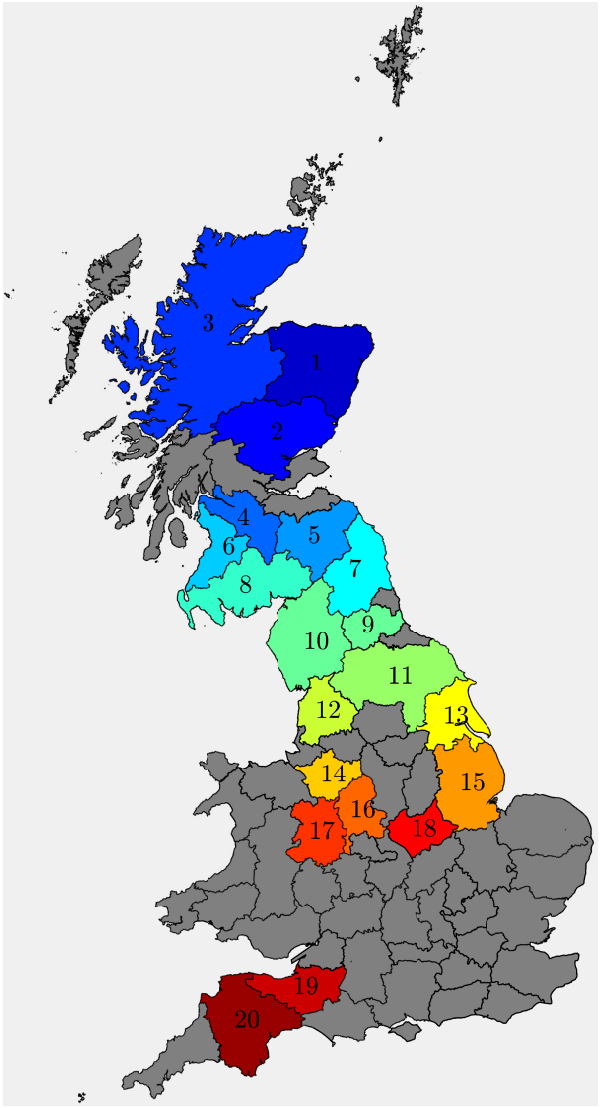

Figure B.19: Map showing the 20 counties with the largest mean number of infected farms after one infectious period (21 days) when epidemics are seeded in Aberdeen and markets are explicitly included. The mean size for each county is given in figure B.20

.

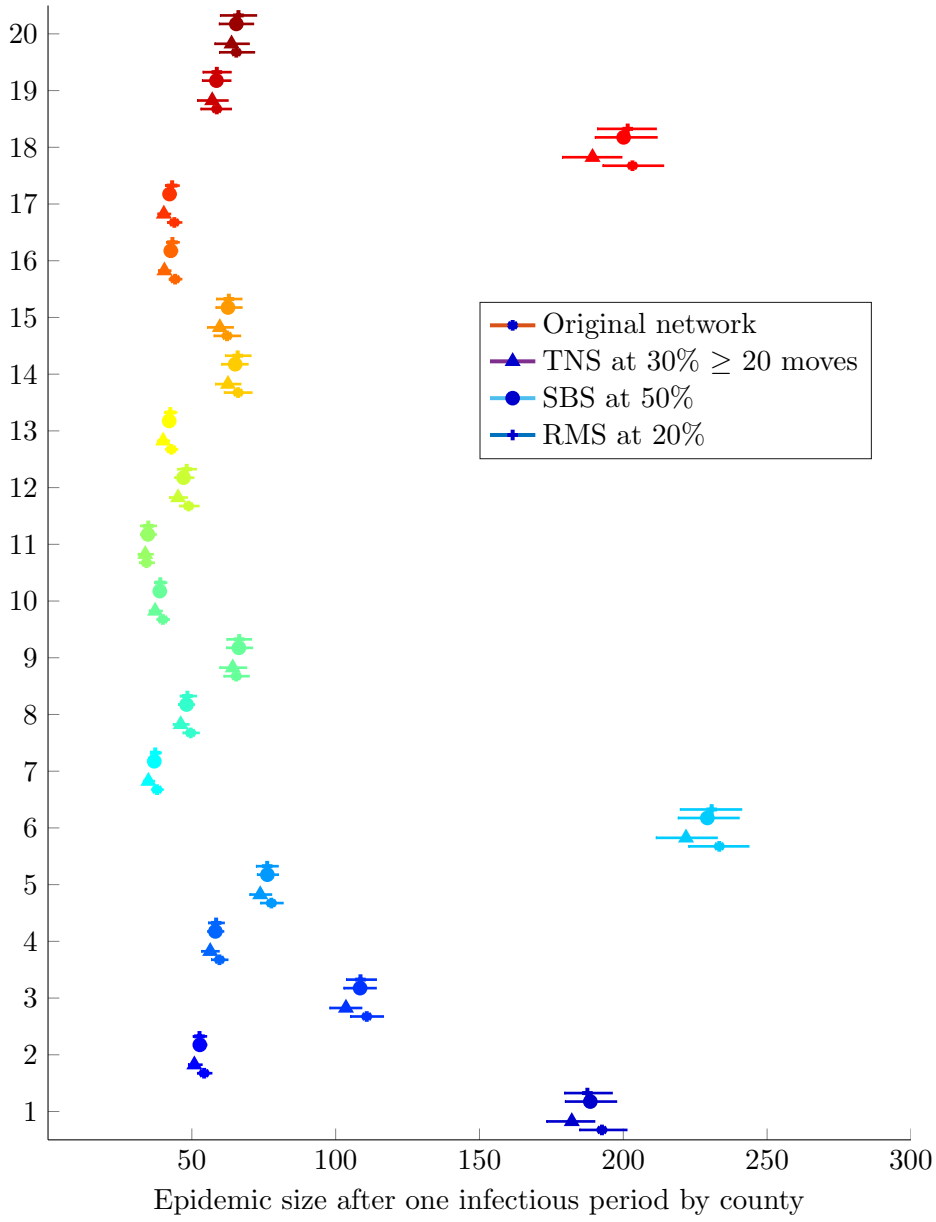

Figure B.20: The average epidemic size for the original network (stars) random movement sampling (RMS) with 10% of sampled movements (crosses), snowball sampling with 50% of nodes (circles) and targeted node sampling (TNS), sampling nodes with more than 20 movements (triangles) for the 20 most infected counties when epidemics are seeded in Aberdeen. Counties are ordered in terms of the proximity of their centroids from Aberdeen with labels corresponding to the map in figure B.19.

Appendix C

Appendix to chapter 4

C.1 The Counties of Great Britain

There are many different geographical divisions of the UK. We have chosen to use one that divides mainland UK into 68 counties. England is comprised of 47 counties which correspond to the 48 ceremonial counties of England but with City of London and Greater London merged together and some other slight differences which are mentioned in table C.1. Scotland has been divided into 14 counties and Wales into 7 comprised of 6 of the 8 preserved counties of Wales with the remaining counties of South, Mid and West Glamorgan merged together. Figure C.1 shows the border and location of all the counties as well as coloured divisions into regions of the UK.

The UK is divided into 12 official region which correspond to the Nomenclature Territorial Units for Statistics (NUTS) level 1, which is a referencing system used by the European Union. We are only interested in mainland UK so can ignore the region of Northern Ireland. We merge the renaming 9 regions of England into 4 giving a total of 6 regions along with Wales and Scotland which are referred to as 'the regions of the UK' throughout this work. The combined regions of England are; North England constituted of North East, North West and Yorkshire and the Humber, the Midlands which combined West Midland and East Midland, East England which is made up of East of England, London and South East, and South West which we keep intact and refer to as South West England.

ID	County	Region	Farms
1	Shetland	Scotland	183
2	Orkney	Scotland	555
3	Eileanan an lar	Scotland	386
4	Highland	Scotland	1844

5	North-East Scotland	Scotland	2427
6	Tayside	Scotland	785
7	Fife	Scotland	299
8	Lothian	Scotland	264
9	Scottish Borders	Scotland	697
10	East Central	Scotland	431
11	Argyll	Scotland	714
12	Clyde Valley	Scotland	867
13	Ayrshire	Scotland	895
14	Dumfries & Galloway	Scotland	1669
15	Northumberland	North England	1063
16	Tyne & Wear	North England	69
17	Durham	North England	894
18	Cleveland and Darlington	North England	232
19	Cumbria	North England	3328
20	Lancashire	North England	1983
21	Cheshire	North England	1628
22	Greater Manchester	North England	424
23	Merseyside	North England	76
24	North Yorkshire	North England	3496
25	East Riding & North Lincolnshire	North England	714
26	West Yorkshire	North England	1058
27	South Yorkshire	North England	432
28	Derbyshire	The Midlands	1703
29	Nottinghamshire	The Midlands	503
30	Lincolnshire excl North	The Midlands	829
31	Leicestershire & Rutland	The Midlands	1049
32	Northamptonshire	The Midlands	593
33	Staffordshire	The Midlands	2021
34	Shropshire	The Midlands	1953
35	Herefordshire	The Midlands	1113
36	Worcestershire	The Midlands	682
37	West Midlands	The Midlands	104
38	Warwickshire	The Midlands	601
39	Norfolk	East England	851
40	Suffolk	East England	452

41	Cambridgeshire	East England	313
42	Bedfordshire	East England	175
43	Hertfordshire	East England	193
44	Essex	East England	369
45	Greater London	East England	50
46	Surrey	East England	298
47	Kent	East England	645
48	East Sussex	East England	584
49	West Sussex	East England	422
50	Hampshire	East England	840
51	Isle of Wight	East England	143
52	Oxfordshire	East England	520
53	Buckinghamshire	East England	539
54	Berkshire	East England	226
55	Gloucestershire excl South	Southwest England	1003
56	North Somerset & South Gloucestershire	Southwest England	748
57	Wiltshire	Southwest England	1092
58	Somerset excl North	Southwest England	2430
59	Dorset	Southwest England	1237
60	Devon	Southwest England	4446
61	Cornwall & The Isles of Scilly	Southwest England	2635
62	North-East Wales	Wales	1739
63	Powys	Wales	2358
64	South Wales	Wales	1514
65	North-West Wales	Wales	1727
66	Ceredigion	Wales	1136
67	Pembrokeshire	Wales	1153
68	Carmarthenshire	Wales	1881

Table C.1: The counties of GB as we have defined them in the text. The ID number refers to figure C.1. The region each county is associated with is defined in the text. We also list the number of farms that had a cattle movement in 2010 for each county.

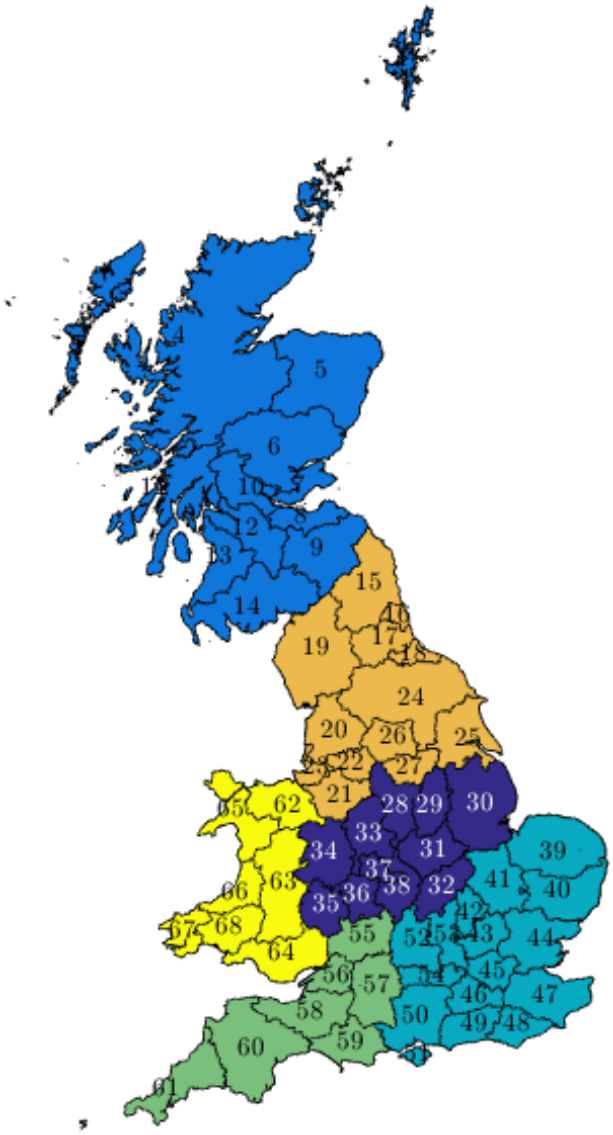

Figure C.1: The counties of the UK divided into 6 regions as described in the text.

C.2 The Provinces of Turkey

Turkey has 81 provinces. Each of these province has a unique number which is used on vehicle licence plates. Like the UK Turkey is divided into 12 official NUTS level 1 regions. We will merge some of the regions to form 7 regions. We merge East Marmara, West Marmara and Istanbul to create Marmara. West Anatolia is merged into Central Anatolia. West Black Sea and East Black Sea become Black Sea. Northeast Anatolia and Central East Anatolia are merged to form Eastern Anatolia. Southeastern Anatolia, Mediterranean, Aegean remain as defined by NUTS.

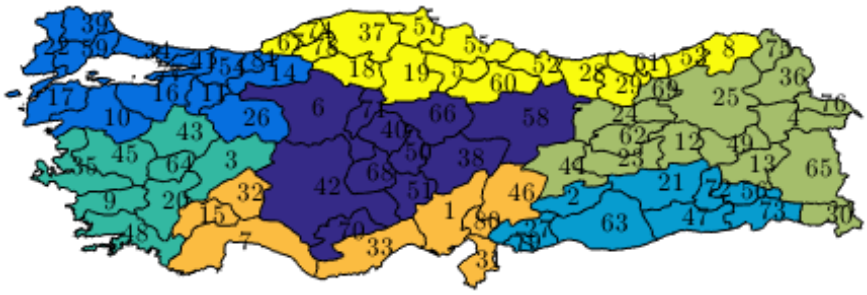

Figure C.2: The provinces of Turkey divided into 7 regions as described in the text.

ID	County	Region
1	ADANA	Mediteranean
2	ADIYAMAN	Southeastern Anatolia
3	AFYON	Aegean
4	AGRI	Eastern Anatolia
5	SANLIURFA	Black Sea
6	AKSARAY	Central Anatolia
7	AMASYA	Mediteranean
8	ANKARA	Black Sea
9	TRABZON	Aegean
10	ANTALYA	Marmara
11	ARDAHAN	Marmara
12	ARTVIN	Eastern Anatolia
13	TOKAT	Eastern Anatolia
14	SIVAS	Marmara
15	SIIRT	Mediteranean
16	AYDIN	Marmara
17	BALIKESIR	Marmara

18	BARTIN	Black Sea
19	BATMAN	Black Sea
20	BAYBURT	Aegean
21	BILECIK	Southeastern Anatolia
22	BINGOL	Marmara
23	BITLIS	Eastern Anatolia
24	BOLU	Eastern Anatolia
25	BURDUR	Eastern Anatolia
26	BURSA	Marmara
27	ZONGULDAK	Southeastern Anatolia
28	CANAKKALE	Black Sea
29	CANKIRI	Black Sea
30	CORUM	Eastern Anatolia
31	DENIZLI	Mediteranean
32	DIYARBAKIR	Mediteranean
33	DUZCE	Mediteranean
34	EDIRNE	Marmara
35	ELAZIG	Aegean
36	ERZINCAN	Eastern Anatolia
37	ERZURUM	Black Sea
38	TUNCELI	Central Anatolia
39	ESKISEHIR	Marmara
40	GIRESUN	Central Anatolia
41	GUMUSHANE	Marmara
42	KASTAMONU	Central Anatolia
43	VAN	Aegean
44	SINOP	Eastern Anatolia
45	HAKKARI	Aegean
46	HATAY	Mediteranean
47	IGDIR	Southeastern Anatolia
48	YALOVA	Aegean
49	SIRNAK	Eastern Anatolia
50	ISPARTA	Central Anatolia
51	ISTANBUL	Central Anatolia
52	IZMIR	Black Sea
53	K-MARAS	Black Sea
54	KARABUK	Marmara

55	KARAMAN	Black Sea
56	KARS	Southeastern Anatolia
57	KAYSERI	Black Sea
58	GAZIANTEP	Central Anatolia
59	KILIS	Marmara
60	KIRIKKALE	Black Sea
61	KIRKLARELI	Black Sea
62	KIRSEHIR	Eastern Anatolia
63	KOCAELI	Southeastern Anatolia
64	YOZGAT	Aegean
65	KONYA	Eastern Anatolia
66	KUTAHYA	Central Anatolia
67	MALATYA	Black Sea
68	ORDU	Central Anatolia
69	MANISA	Eastern Anatolia
70	MARDIN	Central Anatolia
71	TEKIRDAG	Central Anatolia
72	MERSIN	Southeastern Anatolia
73	MUGLA	Southeastern Anatolia
74	MUS	Black Sea
75	NEVSEHIR	Eastern Anatolia
76	NIGDE	Eastern Anatolia
77	OSMANIYE	Marmara
78	RIZE	Black Sea
79	USAK	Southeastern Anatolia
80	SAKARYA	Mediteranean
81	SAMSUN	Marmara

Table C.2: The provinces of Turkey identified by the their licence plate registration number and the region we associate with them.

C.3 Community detection figures

In order to make figure 4.11, figure 4.22 and figure 4.22 which depict the community structures of the UK and Turkey not as points but as areas we developed a protocol to plot high density regions as polygons. Keys steps are depicted in figure C.3 for a community from the Turkey.

First the distance between every node in the community is calculated. Any pairs of farms which are more than a distance x from each other are removed. A binary network is then created from the remaining pairs of farms. The connected components of this network are then found. Components of size 3 or less are ignored. Figure C.3 A depicts all of the farms in the community under considerations while figure C.3 B shows the four remaining components after small components are removed.

For each of these components a Delaunay triangulation is performed on the coordinates of the farms in the component. The Delaunay triangulation for a set of 2D points is the triangulation that guarantees no point is in the circumcircle of any triangle formed by the set of points. The circumcircles for the triangles remaining in the triangulation are stored. Figure C.3 C shows the Delaunay traingulation for one of the components.

At this stage the points on the boundary of the triangulation can be directly used to define the area of the component, this is the convex hull of the triangulation. However, the distribution of points within the convex hull can be quite irregular leading to large areas near the boundary where there are no points. To create an area that is more representative of the actual spatial distribution of the points another distance cut-off parameter y is introduced.

Triangles from the triangulation that have a circumcircle radius less than y are removed. Removing these triangles reduces the number of long range connections along the boundary of the component. The final boundary is now defined using the remaining triangles. Figure C.3 D shows the same component after the circumcircles with large radii have been removed. The difference in the final polygons for all the components after the are final step are shown in figures C.3 E and F.

In creating these figures and those in the main text the cutoff parameters were set at $x = 15$km and $y = 50$km. These values were chosen as they appeared to offer the most detail without fracturing the communities into many components.

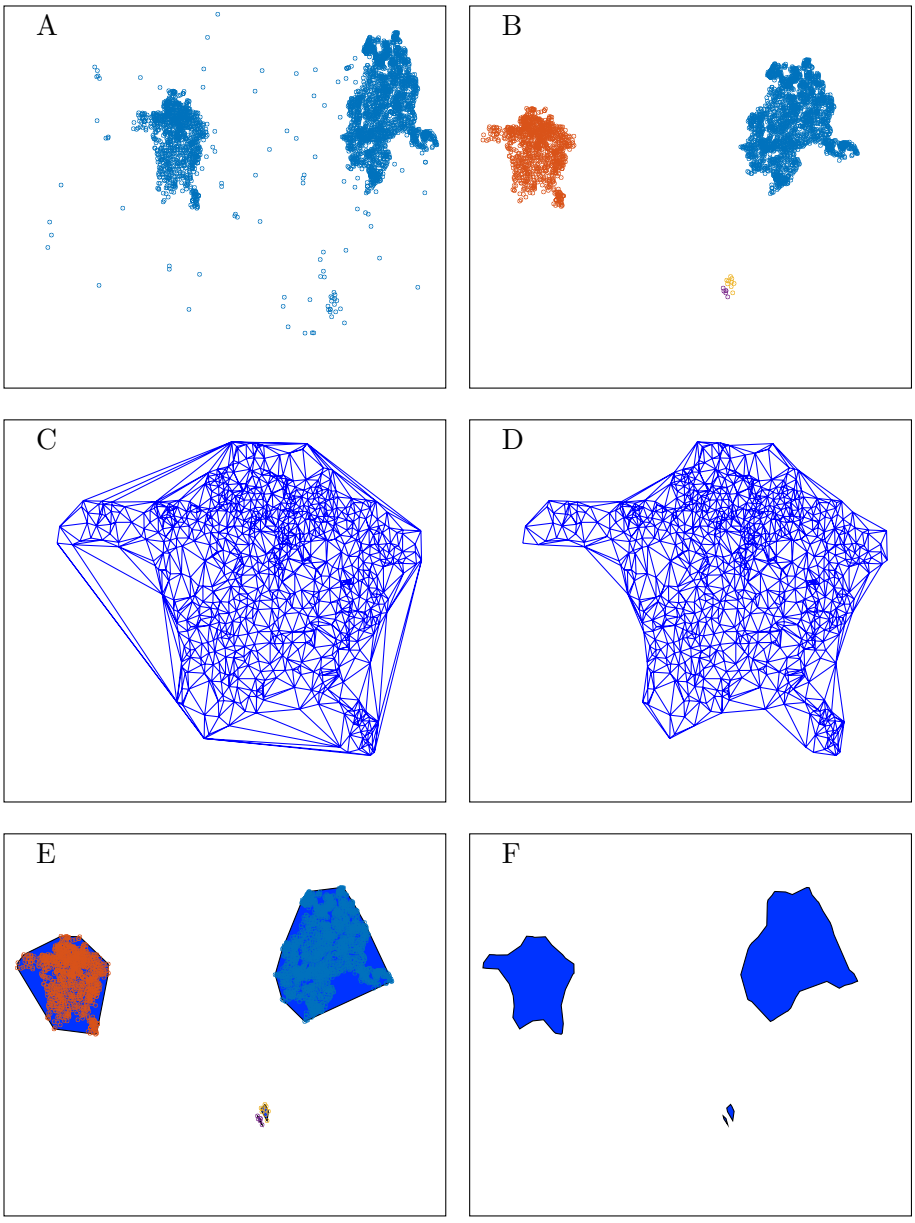

Figure C.3: A walkthrough depicting the main steps in creating figures 4.11, 4.22 and 4.22 using a community from 4.22 A as an example. Each step is explained in detail in the text.

C.4 Further sensitivity analysis for simulations seeded in Devon and Aberdeen

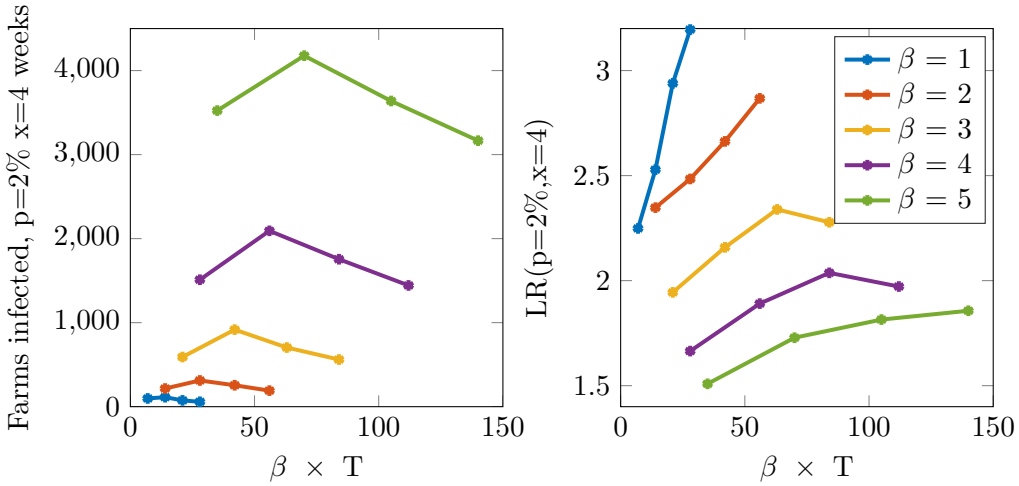

Figure C.4: The effect of varying disease parameters on epidemic size (left) and positive likelihood ratio (right). All analysis takes place 4 weeks into the epidemic or $\beta \in \{1, 2, 3, 4, 5\}$ per day and $T \in \{7, 14, 21, 28\}$ days. We analyse simulations seeded in Devon examining farms that have been infected in at least 2% of simulation.

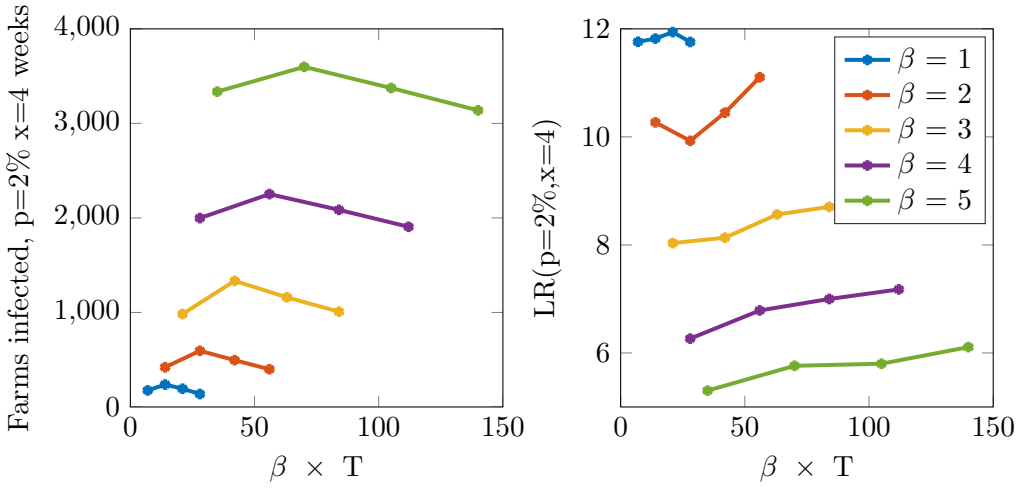

Figure C.5: The effect of varying disease parameters on epidemic size (left) and positive likelihood ratio (right). All analysis takes place 4 weeks into the epidemic or $\beta \in \{1, 2, 3, 4, 5\}$ per day and $T \in \{7, 14, 21, 28\}$ days. We analyse simulations seeded in Aberdeen examining farms that have been infected in at least 2% of simulation.

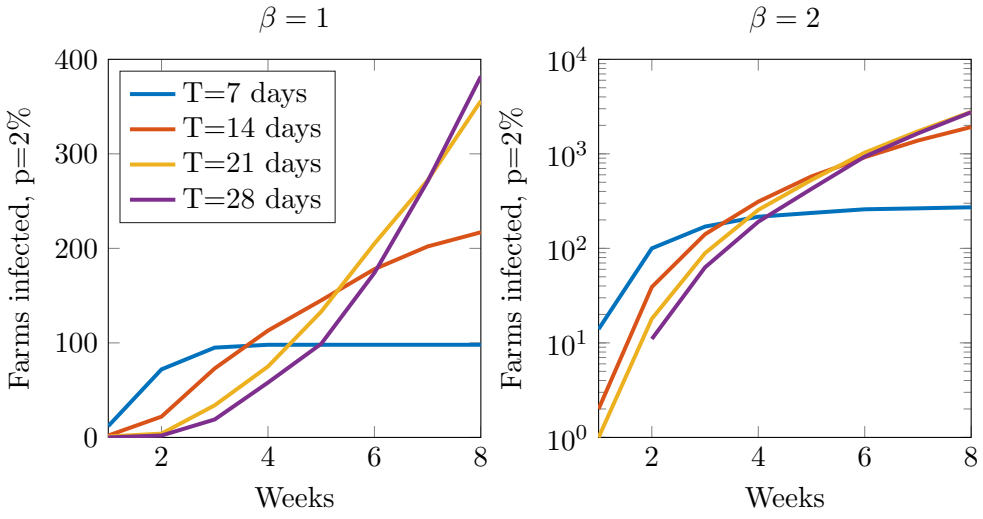

Figure C.6: Plot showing the affect that varying the infectious period T has on the number of infected farms in the early stages of outbreaks when a take-off limit is imposed. On the left we have the transmission parameter $\beta = 1$ and on the right $\beta = 2$. Simulations are seeded in Cumbria and we analyse the farms infected in at least 2% of outbreaks.

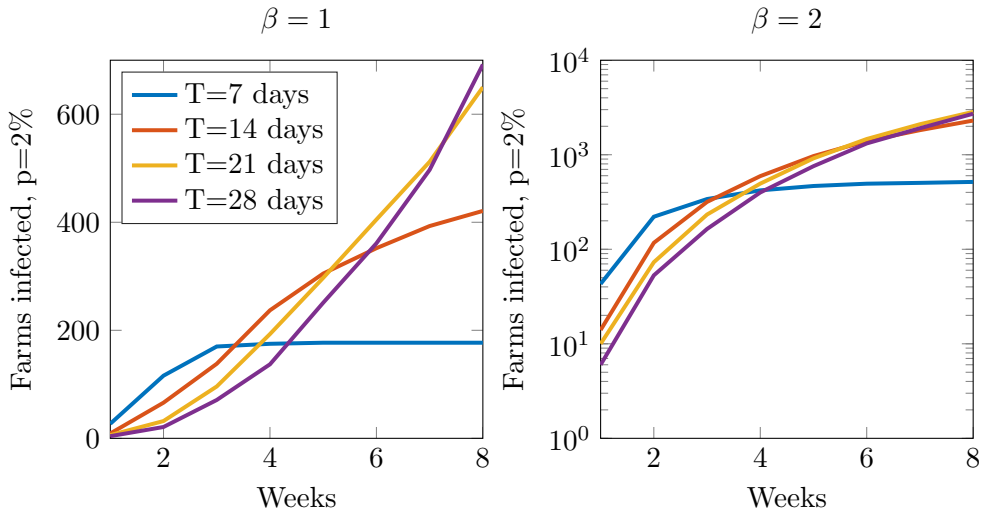

Figure C.7: Plot showing the affect that varying the infectious period T has on the number of infected farms in the early stages of outbreaks when a take-off limit is imposed. On the left we have the transmission parameter $\beta = 1$ and on the right $\beta = 2$. Simulations are seeded in Cumbria and we analyse the farms infected in at least 2% of outbreaks.

Appendix D

Appendix to chapter 5

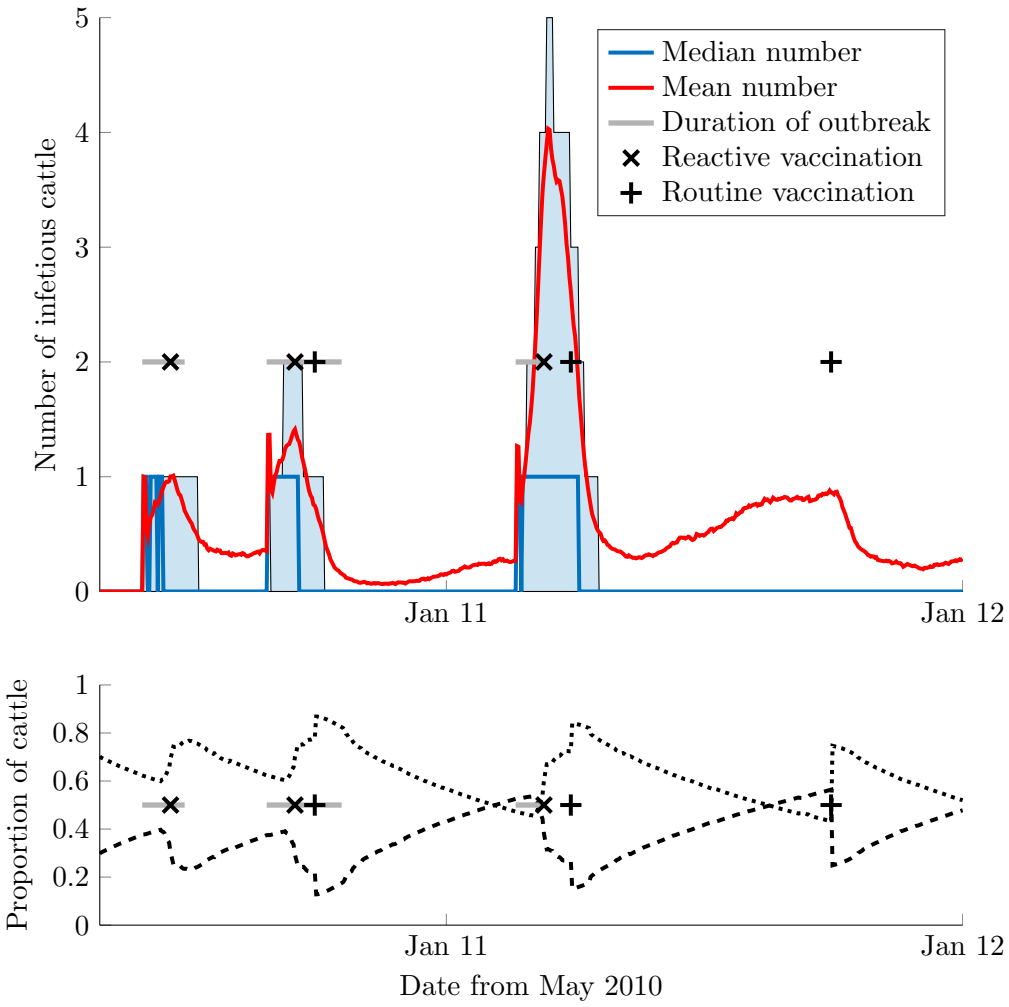

Figure D.1: Simulation results for farm 25,372 using the distribution of parameter values found using ABC with tolerance $\epsilon = 0.60$. The mean and median number of infectious cattle are plotted each day in the upper figure with the interquartile range shaded in grey. The duration of the outbreaks are also shown along with the dates of reactive and routine vaccination. The lower figure shows the proportion of immune (dotted line) and susceptible (dashed line). Only simulations were all three farms had an outbreak take off are included (3,082/10,000)

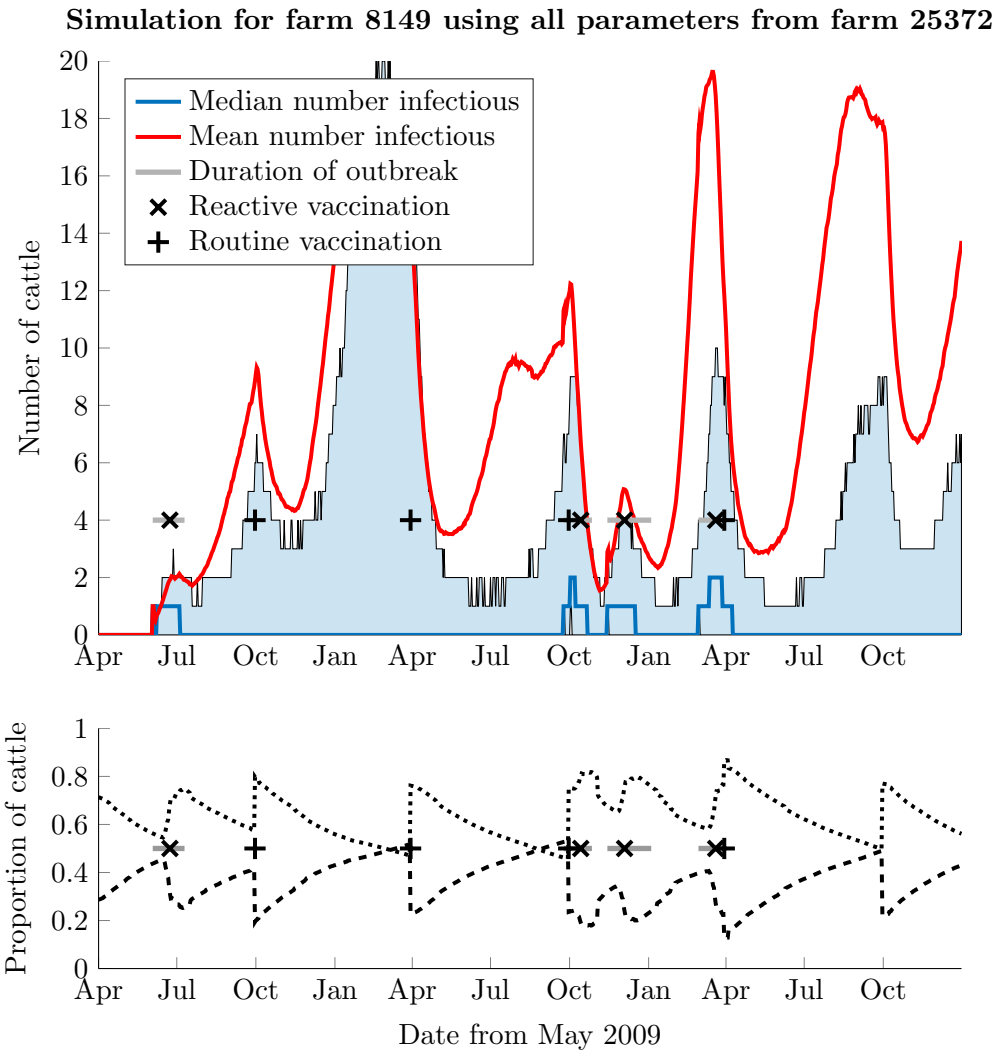

Figure D.2: Simulation results for farm 8,149 with the original vaccination schedule. The median parameters derived from the ABC scheme applied to farm 25,372 with a tolerance of $\epsilon = 0.60$ were used for the simulations. The mean and median number of infectious cattle are plotted each day in the upper figure. The duration of each outbreak is shown as are the dates of the reactive and routine vaccinations. The lower figure shows the proportion of immune (dotted line) and susceptible (dashed line) cattle on the farm. Only simulations were all four outbreaks took off are included (3,936/10,000).

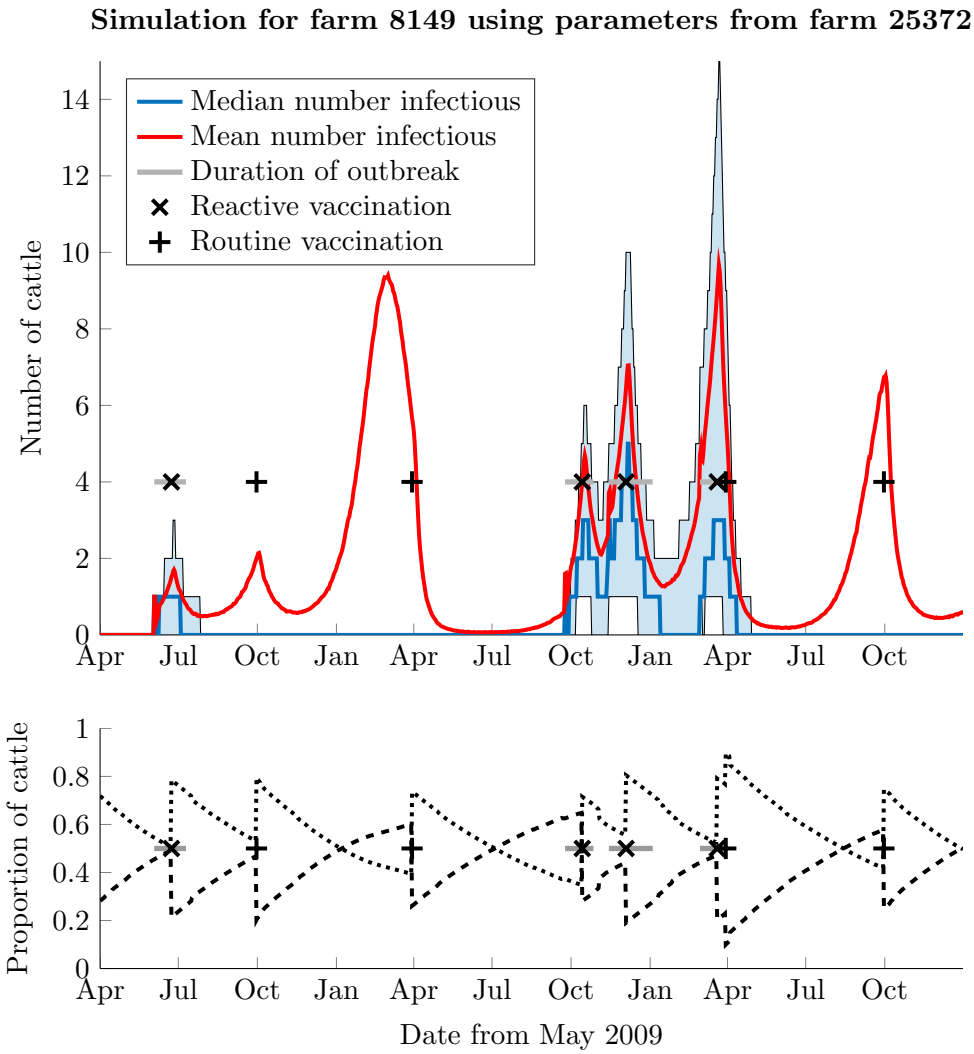

Figure D.3: Simulation results for farm 8,149 with the altered vaccination schedule. The median parameters derived from the ABC scheme applied to farm 25,372 with a tolerance of $\epsilon = 0.60$ were used for the simulations. The mean and median number of infectious cattle are plotted each day in the upper figure. The duration of each outbreak is shown as are the dates of the reactive and routine vaccinations. The lower figure shows the proportion of immune (dotted line) and susceptible (dashed line) cattle on the farm. Only simulations were all four outbreaks took off are included (4,540/10,000).